

Open Research Online

The Open University's repository of research publications and other research outputs

Understanding Emotions in Online Learning: Using Emotional Design and Emotional Measurement to Unpack Complex Emotions During Collaborative Learning

Thesis

How to cite:

Hillaire, Garron (2021). Understanding Emotions in Online Learning: Using Emotional Design and Emotional Measurement to Unpack Complex Emotions During Collaborative Learning. PhD thesis The Open University.

For guidance on citations see [FAQs](#).

© 2019 Garron Edward Hillaire



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Version: Version of Record

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.21954/ou.ro.00012ded>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

UNDERSTANDING EMOTIONS IN ONLINE LEARNING:

USING EMOTIONAL DESIGN AND EMOTIONAL MEASUREMENT TO
UNPACK COMPLEX EMOTIONS DURING COLLABORATIVE LEARNING

GARRON EDWARD HILLAIRE

Thesis Submitted to The Open University
for the degree of Doctor of Philosophy

Institute of Educational Technology

The Open University

1 June 2021

ABSTRACT

Many educational researchers explore the role of emotion in learning and there are many new affordances for emotional measurement. Just as there are many options for emotional measurement there are many theories of emotion. When it comes to the measure of sentiment analysis recent findings suggest it is beneficial to online and blended learning research. The sentiment analysis technologies used for educational research are general purpose technologies suggesting that creating a measure designed for the context of learning would improve the alignment between the measure and context. In addition to aligning measure with the context, there is a need to consider how sentiment analysis relates to emotion theory to determine an appropriate method to evaluate the accuracy of sentiment analysis.

In this PhD thesis I adopt the Constructed Theory of Emotion, which considers emotion as a collective intentionality indicating that consensus on emotion is the best approach toward examining accuracy. From this perspective I create a sentiment analysis measure in the context of learning to contribute to emotional learning analytics the emerging sub-field of learning analytics. The field of learning analytics acknowledges that design and measurement are intertwined. I adopt a design-based research approach by designing supports for emotional communication and examining how such a design impacts the accuracy of sentiment analysis. I then examine correlation analysis with other established measures of emotion. The results contribute to the field of emotional learning analytics by:

- demonstrating promise for generating a classifier based on student perception
- demonstrating benefits of supporting emotion expression in text for students
- demonstrating that students' emotion expression in text does not appear to align with their internal emotional experiences

These findings provide opportunities for further research and suggest caution should be used when interpreting sentiment analysis results in the context of learning.

ACKNOWLEDGEMENTS

There are so many factors that led me to pursuing a PhD making it nearly impossible to acknowledge every aspect that led to applying for a PhD program through to submitting my thesis. Out of everything that has influenced me the one clear thing is that none of this would have happened without my wife and partner Marie Harris. Without Marie influencing my life I would have never enrolled in college and started down an academic path that has ultimately led me to this PhD program. For Marie, I would like to express my gratitude in changing my life in so many ways that I would never have considered including connecting me to the academic world. She is of course the smartest person I know and there is clear evidence of this as she was intelligent enough to *not* pursue a PhD. This circumstance made my decision to pursue one a decision she personally would have avoided. I have to express my gratitude again to Marie for allowing me to change her life in a way she would not have considered. It has cost us both so much to get to this point and I can only continue to apologetically express gratitude as I pursue the precarious and fragile path of an academic career.

Both of our lives were changed by my pursuit of a PhD in part by shaping our family life as we raised our two sons Helo and Xander. I am grateful for the two lovely young men we have in this family as they provided an insightful and honest perspective for our family in this endeavor. Near the end of the process my sons would frequently ask me: “is the research done?” I would usually respond by saying that research is never done, but that I was making progress on my thesis. I suspect my sons get their intelligence from their mother as they have both expressed that they do not plan to do research when they grow up.

While my immediate family went through this process with me day by day and long night after long night there are even more family member who I would like to acknowledge. My mother Diane Hillaire who has been unapologetically supportive in everything I have ever done. In fact, as I took my family to the UK all of the grandmothers and grandfathers of my children are owed a debt of gratitude for their love and support over the past few years. Thank you to grandma Marge Saffer, grandparents Linda & Ramsey Younger, and grandparents Andy & Jean Harris. I would also like to thank my sister Tawnya El-Masry and brother Voir Hillaire.

In jest I make fun of my decision to go down this road, but the reality is that it is just simply a crazy thing to pursue. I have effectively spent the last five years considering what happened during a one-hour session of a classroom in the Netherlands. At the end of five years I feel comfortable saying that there are some things that I understand but I have even more questions. I have many people to thank in terms of making the pursuit of my crazy ideas possible.

I have to thank first and foremost my supervisor Bart Rienties for rarely agreeing with me but always being supportive in spite of our differences of opinion. I cannot think of a single person whom I have had as many disagreements with and yet as I complete my thesis work I can see how working with you for the past four years has resulted in a higher quality of work and developed my perspective as a researcher. As I conducted research my fascination with emotions and learning developed countless off-shoot questions that would gain my attention (what Bart would refer to as my crazy ideas). As I would excitedly share my new fascinations they were frequently met with “So what?” and “Who cares?” I can comfortably say at this stage that it is you Bart who cares at least enough to patiently help me to gain focus in my work, rigor in my methods, and clarity in my expression. In addition, my supervisor Mark Fenton-O’Creevey who at times felt like the adult in the room who could referee arguments between Bart and myself. I believe you provided a great deal of insight on emotions and I have, I believe, demonstrated a reasonable understanding in your eyes with the work I submit for this PhD thesis. My third supervisor, Zdenek Zdrahal, also provided support early in the process and I thank you for all that you did in getting me on track towards my dissertation.

While my supervision team was instrumental the entire community at The Open University provided me with the needed support as I undertook this work. In terms of staff I would like to thank Rebecca Ferguson, Doug Clow, Elizabeth Fitzgerald, Thomas Ulman, and Jekaterina Rogaten. In terms of students I would like to thank Janesh Sanzgiri, Jenna Mittelmeier, Ralph Mercer, Francisco “Paco” Inniesto. Each of you helped me in times of need throughout this process.

DECLARATION OF AUTHORSHIP

The work detailed in this PhD thesis is comprised of two experiments. The first experiment was done in conjunction with Jenna Mittelmeier. While I collaborated on the design of the experiment Jenna's focus was on the content of the activity and how the activity influenced participation. I focused on how students perceived emotion in their own communication when reflecting on their participation in the activity. This clear distinction kept our parallel efforts unique and original contributions. In addition, I worked with Dirk Temeplaar the teacher at the site of the both the 2016 and 2017 experiment and while his input was tremendously valuable his feedback focused on modifications needed to make my research designs appropriate for the setting. I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. Any use of the works of any other author is acknowledged at their point of use.

LIST OF ABBREVIATIONS

<ul style="list-style-type: none"> • ABS – Absolute Value • AC – Affective Computing • AL – Affective Learning • BEQ – Berkley Expressivity Questionnaire • BET – Basic Emotion Theory • CPM – Component Process Model • CSCL – Computer Supported Collaborative Learning • CTE – Constructed Theory of Emotion • DBR – Design Based Research • ELA – Emotional Learning Analytics • EM – Expectation Maximization Algorithm • ESM – Evaluative Space Model • ESS – Emotional Sentence Starters • FN – False Negative • FP – False Positive • LA – Learning Analytics • LD - Learning Design • LIWC - Linguistic Inquiry and Word Count 	<ul style="list-style-type: none"> • MES – Mixed Emotion Scale • MCB – Majority Class Baseline • MIN – Minimum Value • MTURK – Mechanical Turk • NB – Naïve Bayes • PANAS – Positive Affect Negative Affect Schedule • PBL – Problem Based Learning • PBR – Polarity Bias Rate • RB – Random Baseline • SA – Sentiment Analysis • SAT – Situated Affectivity Theory • SRL – Self Regulated Learning • SSRL – Self and Socially Regulated Learning • SSSAC – Student Sourced Sentiment Analysis Classifier • SVM – Support Vector Machine • TN – True Negative • TP – True Positive • UDL – Universal Design for Learning • UME – Univariate Mixed Emotion
---	--

TABLE OF CONTENTS

Abstract	3
Acknowledgements.....	4
Declaration of Authorship.....	6
List of Abbreviations	7
List of Figures	14
List of Tables	15
Chapter 1 Introduction.....	16
1.1 Background.....	16
1.2 Problem Definition	18
1.3 Proposed Solution	22
1.4 Thesis Structure.....	28
1.4.1 chapters	28
1.4.1.1 chapter 2 literature review	28
1.4.1.2 chapter 3 methodology	29
1.4.1.3 chapter 4 study 1.....	29
1.4.1.4 chapter 5 study 2.....	30
1.4.1.5 chapter 6 study 3.....	32
1.4.1.6 chapter 7 conclusions.....	32
Chapter 2 Literature Review	34
2.1 Theory and Measurement of Emotion.....	34
2.1.1 component process model of emotion and communication	37
2.1.2 what is sentiment analysis?	41
2.1.2.1 mapping emotional theory to sentiment analysis studies in the context of learning	43
2.1.2.2 implications from emotion theory for emotion learning analytics	46
2.1.3 how sentiment analysis measures emotion using valence	47
2.1.3.1 what is valence?	47
2.1.3.1.1 bipolar perspective on valence.....	48
2.1.3.1.2 bivariate perspective on valence	51
2.1.3.1.3 valence as both bipolar and bivariate	55
2.1.3.1.4 bivariate, univariate, and multivariate mixed paradigms.....	56
2.1.3.2 sentiment analysis methods	57
2.1.3.2.1 ground truth and sentiment analysis	57
2.1.3.2.2 ground truth used in This thesis	61
2.1.3.2.3 sentiment analysis methods.....	61
2.1.3.2.4 sentiment analysis and domain dependency	62
2.1.3.2.5 mapping valence theory to sentiment analysis studies in the domain of learning..	63
2.1.4 emotional measures beyond verbalization and communication.....	66
2.1.4.1 state.....	66
2.1.4.2 trait.....	69
2.2 Theory and Design of Learning	70
2.2.1 learning theory and emotion.....	70

2.2.2 learning design and emotion	72
2.2.2.1 computer supported collaborative learning (CSCL)	72
2.2.2.2 universal design for learning (UDL)	73
2.2.2.3 scripting supports informed by CSCL and UDL	73
2.3 Conclusion.....	75
Chapter 3 Methodology.....	78
3.1 Introduction	78
3.2 Ontology and Epistemology	78
3.3 Overview of Adopted Approach.....	82
3.3.1 design based research and experimental design	82
3.3.2 three studies to test the six research questions.....	83
3.4 Instruments Used	86
3.4.1 pre-questionnaires.....	86
3.4.1.1 berkley expressivity questionnaire (BEQ)	86
3.4.2 data collected during the computer lab	87
3.4.2.1 udio.....	87
3.4.2.1.1 react	87
3.4.2.1.2 discuss It.....	88
3.4.2.1.3 iterating the design of udio with emotional sentence starters.....	89
3.4.3 data collected at the end of the computer lab	90
3.4.3.1 PANAS.....	90
3.4.3.2 mixed emotion scale (MES).....	91
3.4.4 data collected after the computer lab	91
3.4.4.1 post-survey	91
3.4.4.2 semi-structured interviews	92
3.4.5 data and sentiment analysis tools	94
3.4.5.1 SentiStrength	94
3.4.5.2 VADER.....	94
3.4.5.3 séance.....	95
3.4.5.4 LIWC	95
3.4.5.5 majority class baseline (MCB)	96
3.4.5.6 random baseline (RB)	96
3.4.6 machine learning classifiers.....	96
3.4.6.1 pre-processing text.....	97
3.4.6.2 logisitc regression.....	98
3.5 Data analysis	99
3.5.1 annotated data	99
3.5.1.1 student sourced labels.....	99
3.5.1.2 mechanical turk labels	100
3.5.1.3 collective intentionality, ground truth, and the expectation maximization algorithm ..	101
3.5.1.4 examing the effects of emotional sentence starters.....	101
3.5.2 reliability	102
3.5.3 evaluating the classifier.....	102
3.5.3.1 cross-valdiation	102
3.5.3.2 testing classifiers on novel data	104
3.5.3.3 student interviews.....	104
3.5.4 correlaion analysis	105
3.6 Ethics.....	106

3.7 Conclusion	107
Chapter 4 Study 1 – Introducing Student Sourced Sentiment Analysis.....	108
4.1 Introduction.....	108
4.2 Methods	112
4.2.1 setting	113
4.2.2 2016 procedure and participants	113
4.2.3 2017 procedure and participants	115
4.2.4 instruments	117
4.2.4.1 student sourced examples.....	117
4.2.4.2 heuristic sentiment analysis instruments	117
4.2.4.3 lexical sentiment analysis instruments.....	118
4.2.4.4 machine learning sentiment analysis instrument	120
4.2.5 analysis.....	120
4.2.5.1 crowd sourced labels.....	120
4.2.5.2 pre-processing Text.....	120
4.2.5.3 processing text.....	121
4.2.5.4 ten-fold cross-validation.....	122
4.2.5.5 comparative analysis method	123
4.3 Results	124
4.3.1 data collection and ground truth established using the expectation maximization algorithm for three datasets	124
4.3.1.1 2016 data collection (training data).....	124
4.3.1.2 2016M mechanical turk data collection (training data).....	127
4.3.1.3 2017C data collection (test data).....	128
4.3.2 RQ1: to what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters?	129
4.3.2.1 RQ1A: to what extent do students agree in terms of inter-rater agreement when providing examples?	129
4.3.2.2 RQ1B: To what extent do mechanical turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples?.....	130
4.3.2.3 Answering RQ1: to what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters?.....	131
4.3.3 RQ2: to what extent can crowd sourced, and in particular student sourced, examples train a machine learning classifier to predict the valence categories of positive, negative, neutral, and mixed?	131
4.3.3.1 RQ2A: to what extent can student labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed?.....	132
4.3.3.2 RQ2B: to what extent can Mechanical Turk labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed?.....	133
4.3.3.3 RQ2C: how do logistics classifiers trained using student labels and Mechanical Turk labels compare to general benchmarks when predicting the valence categories of positive, negative, neutral, and mixed?	134
4.3.3.4 RQ2D: to what extent do students find predictions from a student sourced classifier useful?.....	136
4.3.3.5 Answering RQ2 to what extent can crowd sourced, and in particular student sourced, examples train a machine learning Classifier?	138
4.3.4 sample conversation, student labels, and predictions by SSSAC logistic and SentiStrength. 138	
4.4 Discussion	146
4.5 Limitations and Future Research	147

Chapter 5 Study 2 – Improving a Student Sourced Sentiment Analysis Classifier With Emotional Design Using Emotional Sentence Starters - A Randomized Control Trial 149

5.1 Introduction	149
5.2 Methods.....	151
5.2.1 setting	151
5.2.2 procedure.....	152
5.2.3 participants.....	154
5.2.4 instruments	159
5.2.4.1 student sourced examples.....	159
5.2.4.2 student sourced sentiment analysis instruments.....	159
5.2.5 analysis.....	159
5.2.5.1 comparative analysis	159
5.3 Results	160
5.3.1 student sourced labels	160
5.3.2 RQ3: to what extent can emotional sentence starters improve the inter-rater reliability of student examples?	162
5.3.3 RQ4: to what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?.....	163
5.3.3.1 Answering RQ4: To what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?	167
5.3.4 sample conversation from 2017S dataset, student labels, and predictions by SSSAC logistic(2016) and SSSAC logistic(2017S)	167
5.4 Discussion	171
5.5 Limitations and Future Work	172

Chapter 6 Study 3 – Exploring the emotional journey of students from Dispositions, Incidental Emotions, Emotional Expression, and Overall Emotional Experience 173

6.1 Introduction	173
6.2 Methods.....	176
6.2.1 setting	176
6.2.2 procedure.....	176
6.2.3 participants.....	177
6.2.4 instruments	177
6.2.4.1 berkley expressivity questionnaire (BEQ)	177
6.2.4.2 react.....	183
6.2.4.3 gag-of-words ensemble valence (SSSAC Logistic).....	185
6.2.4.4 mixed emotion scale (MES).....	185
6.2.4.5 PANAS.....	187
6.3 Analysis	190
6.4 Results	191
6.4.1 RQ5: to what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?.....	191
6.4.2 RQ6: to what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?	196

6.5 Conclusions.....	198
6.6 Limitations and Future Work.....	200
Chapter 7 Conclusion.....	202
7.1 Introduction.....	202
7.2 Findings	202
7.3 Unique Contribution.....	206
7.4 Limitations.....	208
7.5 Recommendations.....	210
7.5.1 future work on sentiment analysis for valid self-report of emotion.....	211
7.5.2 future work on sentiment analysis for regulated communication	211
7.5.3 future work on sentiment analysis and categorized communication	212
7.6 Conclusion	214
References.....	215
Appendices.....	231
Appendix 1 – BEQ.....	231
1. The entire tool.....	231
2. items administered.....	231
Appendix 2 – PANAS.....	232
1. The entire tool administered	232
Appendix 3 – MES.....	232
1. The entire tool.....	232
2. items administered.....	233
Appendix 4 – Post Activity for Experiment 1	233
Section 1 - Reflection and recall of group work process.....	234
Section 2 - Reflection of assignment content	235
Section 3 - Evaluation of individual contributions.....	237
Section 4 - Emotional reactions to data.....	239
Section 5 - Soft skills and supports.....	243
Section 6 - Feedback to the members of my lab group.....	244
Appendix 5 - Post Activity For Experiment 2	247
Section 1 - Reflection and recall of group work process.....	247
Section 2 - Reflection of assignment content	248
Section 4 - Emotional reactions to data.....	251
Section 5 - Soft skills and supports.....	254
Section 6 - Feedback to the members of my lab group.....	255
Appendix 6 – HREC for 2016	258
1. Title of project.....	258
2. Abstract.....	259
3. Investigators.....	259
4. Literature review	260
5. Methodology.....	262
6. Participants	264
7. Recruitment procedures	265

8. Consent.....	265
9. Location(s) of data collection	265
10. Schedule.....	266
11. Published ethics and legal guidelines to be followed	267
12. Data protection and information security.....	267
13. Research Data Management	267
14. Deception.....	268
15. Risk of harm.....	268
16. Debriefing.....	268
17. Research organisation and Funding.....	269
18. Other project-related risks.....	269
19. Benefits and knowledge transfer.....	269
20. Disseminating and publishing research outcomes	269
21. Declaration.....	270
Appendix.....	271
Appendix 7 – Permission Form for 2016	271
Appendix 8 – HREC for 2017.....	273
1. Title of project	274
2. Abstract.....	274
3. Investigators.....	275
4. Literature review	275
5. Methodology.....	277
6. Participants	280
7. Recruitment procedures	281
8. Consent.....	281
9. Location(s) of data collection	281
10. Schedule.....	283
11. Published ethics and legal guidelines to be followed	283
12. Data protection and information security.....	283
13. Research Data Management	283
14. Deception.....	284
15. Risk of harm.....	284
16. Debriefing.....	284
17. Research organisation and Funding.....	285
18. Other project-related risks.....	285
19. Benefits and knowledge transfer.....	285
20. Disseminating and publishing research outcomes	285
21. Declaration.....	286
Appendix.....	287
Appendix 9 – Permission Form for 2017	287
Appendix 10 – Interview Protocol for 2017.....	289
Appendix 11 - Addendum	291

LIST OF FIGURES

1 Figure 2.1 Three theoretical perspectives on emotion: Basic, Constructed, and Situated	37
2 Figure 2.1.1 Component Process Model (CPM) of emotion adapted from Scherer (2009).....	39
3 Figure 2.1.3.1 Three Models of Valence: Bipolar, Bivariate, and Evaluative Space...	48
4 Figure 2.1.3.1.1a Precipice Effect for Stimuli - Unnecessary.....	49
5 Figure 2.1.3.1.1b - Bimodality Effect for Stimuli - Completely Indifferent.....	50
6 Figure 2.1.3.1.2 All possible bivariate rating with positive and negative scales from 1-5 categorized	55
7 Figure 3.3.2 Study design of Experimental Study 1 and Study 2	84
8 Figure 3.4.2.1.1 Interface of 'React' to self-report emotional response	88
9 Figure 3.4.2.1.2 Discussion interface with Sentence Starters.....	89
11 Figure 3.5.3 10-Fold Cross-validation.....	103
12 Figure 4.2.2 Pilot Experiment Conducted in 2016	114
13 Figure 4.2.3a - Interface of 'React' to self-report emotional response	116
14 Figure 4.2.3b Input, Process, Output for Main Study conducted in 2017	116
15 Figure 5.2.2a - Interface of 'React' to self-report emotional response	153
16 Figure 5.2.2b - Discussion interface with Sentence Starters.....	154
17 Figure 5.2.2c Input, Process, Output for Main Study conducted in 2017.....	154
18 Figure 5.2.3 - Participation by Research Condition for Main Study conducted in 2017	155
19 Figure 6.2.4.1a – BEQ three constructs compared to one construct for nine items ..	178
20 Figure 6.2.4.1b – BEQ three constructs compared to one construct for six items ...	179
21 Figure 6.2.4.1c – Histogram of BEQ Positive scores	181
22 Figure 6.2.4.1d – Histogram of BEQ Negative scores	182
23 Figure 6.2.4.2 – Histogram of React Scores Reporting Incidental Emotions to the Warm-up Activity.....	184
24 Figure 6.2.4.4a Comparing a one-factor model to a two-factor model of the Mixed Emotion Scale	186
25 Figure 6.2.4.4b – Histogram of MES Mixed Scores	187
26 Figure 6.2.4.5a – Item loading for 12 item PANAS scale comparing one-construct to two-constructs	188
27 Figure 6.2.4.5b – Item loading for 11 items PANAS scale comparing one-construct to two-constructs	189
28 Figure 6.2.4.5c – Histogram of PANAS Positive Scores.....	189
29 Figure 6.2.4.5d – Histogram of PANAS Negative Scores.....	190

LIST OF TABLES

Table 2.1.2.1 Emotional theory inferred by measures of accuracy of sentiment analysis in the context of learning.....	45
Table 2.1.3.1.2 Five bivariate scores interpreted by five methods to compute mixed emotion expression.....	54
Table 2.1.3.2.5 Valence categories for SA used in the context of learning	65
Table 3.3.2 Overview of Study 1, Study 2, and Study 3	86
Table 4.3.1.1a Number of Students Per Example	125
Table 4.3.1.1b Student Examples and Associated Labels	125
Table 4.3.1.1c EM Selection of Student Labels.....	126
Table 4.3.1.2 Student Examples and Associated Mechanical Turk Labels.....	128
Table 4.3.1.3 Ground Truth Labels of chat messages into four emotion categories in 2017C	129
Table 4.3.2.1 Student Sourced Labels Agreement Statistics	130
Table 4.3.2.2 Student Sourced Labels Agreement Statistics	131
Table 4.3.3.1a Accuracy of classifier trained on student labels predicting novel student labels.....	132
Table 4.3.3.1b Accuracy of classifier trained on student labels using 10-fold cross validation	133
Table 4.3.3.2a Accuracy of classifier trained on mechanical turk labels predicting novel student labels	133
Table 4.3.3.2b Accuracy of classifier trained on Mechanical Turk labels using 10-fold cross validation	134
Table 4.3.3.3 Accuracy of Predicting Student Sourced Labels with F-Measures	135
Table 4.3.4a - An Example Conversation.....	140
Table 5.2.3a Age by Research Condition for Main Study conducted in 2017	156
Table 5.2.3b Gender by Research Condition for Main Study conducted in 2017	156
Table 5.2.3c Nationality by Research Condition for Main Study conducted in 2017 ..	157
Table 5.2.3d Berkley Expressivity Questionnaire (BEQ) scores by Research Condition for Main Study conducted in 2017.....	158
Table 5.3.1 Ground Truth Labels of chat messages into four emotion categories in 2016 & 2017 (control and intervention condition)	161
Table 5.3.4 Sample Conversation with Labels from ESS condition.....	168
Table 6.2.4.1a The Cronbach's alpha for the three constructs of the Berkley Expressivity Questionnaire	180
Table 6.2.4.1b Berkley Expressivity Questionnaire items and text retained after exploratory factor analysis.....	180
Table 6.2.4.4a Mixed Emotion Scale (MES) items and text retained after exploratory factor analysis	186
Table 6.4.1 Correlations with confidence intervals for SSSAC Logistic and comparison measures	194
Table 6.4.2 Correlations with confidence intervals for SentiStrength and comparison measures	197

CHAPTER 1 INTRODUCTION

1.1 BACKGROUND

Many researchers have investigated the role of emotion in learning as evidenced by a recent review of new affordances of emotional measurement in learning covering 100 different studies (Rienties & Rivers, 2014). Some educational researchers make strong claims that the whole process of learning is emotional based on evidence that contemporary neuroscience considers decision-making, memories, and actions closely intertwined with emotions (Immordino-Yang & Damasio, 2007). Other researchers draw the conclusion that learning is emotional because failure is a part of learning, and failure is likely to cause emotional reactions (D'Mello, Taylor, & Graesser, 2007). Similarly some educational researchers focuses on emotions associated with achievement (Pekrun, 2005; Pekrun, Goetz, Frenzel, Barchfeld, & Perry, 2011). It is perhaps reasonable to say that there is reason to examine emotions in learning. Furthermore, there is an increased interest in using contemporary measures to better understand the extent to which emotions are either integral to learning or influence learning.

To better understand how measures of emotion may provide insight into learning it is important to consider how such measures relate to theory on emotion (Weidman, Steckler, & Tracy, 2016). For example, when researchers consider emotion in learning they focus on a variety of aspects of emotion which results in a complex set of evidence that speaks to many different facets of emotion (Pekrun, 2005). As there are many potential emotional measures in the context of learning (Rienties & Rivers, 2014) and even more measures used for general emotional research (Weidman et al., 2016) it is important to clarify the target of a measure to identify an effective evaluation strategy. For example, one major disagreement on emotional theory is the hundred year emotion war (Lindquist, Siegel, Quigley, & Barrett, 2012), where some consider emotions best considered in universal discrete terms, like anger and happiness (Ekman et al., 1992;

Tracy & Randles, 2011), while others advocate that our common physiological experiences of emotions are best understood in dimensional terms, such as valence (positive to negative) and arousal (low energy to high energy) (Russell & Barrett, 1999a; Russell & Carroll, 1999). In this thesis I take a dimensional perspective on emotion (Calvo & D'Mello, 2010).

Within dimensional theory there are further debates on how to model valence. Some advocate that valence is bipolar which means it is modeled on one dimension from negative to positive (Russell & Carroll, 1999). Others argue that valence is bi-variate, meaning positive and negative co-activate independently (Watson, Wiese, Vaidya, & Tellegen, 1999). These debates eventually raise questions about the validity of measures used to research emotion as a way to make sense of conflicting results. Investigations that critically examine measures based on opposing theories find that how measures are designed influence what they measure effectively, stating that our theory about emotion influence what we measure (Green, Goldman, & Salovey, 1993). By taking a step back from the conflicting evidence on valence there is a third position that acknowledges emotion is at times bipolar and at times bi-variate (Cacioppo, Gardner, & Berntson, 1999). This third position reaffirms that emotion research at present has a complex set of evidence that speaks to many different facets (Pekrun, 2005). As I explore contributions to our understanding of emotion with the backdrop of these theoretical debates, this sets the tone that I need valid and reliable measures of emotion with clearly articulated theoretical underpinnings to effectively evaluate what those measures can contribute to our theoretical understanding of emotion. The measure I focus on this thesis is Sentiment Analysis (SA) which is commonly defined as the detection of how the opinion of the author of the text elicits a reaction from the intended reader of the text (Balahur & Steinberger, 2009). SA frequently categorizes text in dimensional terms such as determining if text is positive or negative (Calvo & D'Mello, 2010). This is potentially valuable as a measure because it may provide insight into the emotional state of students which could help in terms of potentially gaining insights into psychological processes of learning (Hillaire, Rienties, & Goldowsky, 2018), predicting course evaluations (Rajput, Haider, & Ghani, 2016), and predicting retention rates (Wen, Yang, & Rosé, 2014). One thing that researchers do agree on in terms of SA to detect valence

in text is that this measurement approach is very context sensitive (Pang & Lee, 2006). This indicates that I need to also understand the context of measurement in conjunction with the strengths and limitations of the emotional theory that underpins the measure.

In the context of learning there are many competing theories of what constitutes learning (Edgar, 2012). Learning theories can range from a focus on social aspects of learning (Ferguson & Shum, 2012; Van Den Bossche, Gijssels, Segers, & Kirschner, 2006; Vygotsky, 1978), individual influences including an emphasis on self-regulation during learning (Hadwin, Nesbit, Jamieson-Noel, Code, & Winne, 2007; Perry & Winne, 2006; Winne & Hadwin, 1998), the intersection between the role of self and social aspects for regulation (Järvelä, 2014; Järvelä & Hadwin, 2013), and the influence design has on learning with frameworks such as Learning Design (LD) (Bakharia et al., 2016; Rienties, Nguyen, Holmes, & Reedy, 2017) or Universal Design for Learning (UDL) (Coy, Marino, & Serianni, 2014; Rappolt-Schlichtmann & Daley, 2013; Rose, Harbour, Johnston, Daley, & Abarbanell, 2006). To investigate complex questions of emotional measurement in the context of learning it is therefore also necessary to establish a theory of learning and consider how the design of the learning context provides affordances for measurement (Rienties et al., 2017; Wise & Vytasek, 2017).

1.2 PROBLEM DEFINITION

To make progress toward understanding the role of emotion in learning, there is a need to consider the quality of emotional measures in the context of learning. New affordances of measurement has many researchers exploring how these measures relate to learning (Rienties & Rivers, 2014). There is a more general concern that many researchers are using emotional measures haphazardly (Weidman et al., 2016). For example, a review on the use of emotional measurements in the journal *Emotion* over a ten-year period found that when research is conducted using a stated theoretical perspective on emotion there is a lack of consistency between theory and measures used by researchers (Weidman et al., 2016). Similar observations have been made that when researchers measure emotions such as anger and fear, they sometimes recode their measures into the dimension of valence (positive to negative) considering anger and fear to both be negative (Calvo & D'Mello, 2010). While this may seem like a subtle move,

the theory behind discrete emotions (e.g., anger, fear, frustrations) is most prominently rooted in a universal perspective on emotion holding that everyone has a common experience of some emotions, referred to as basic emotions, like anger (Tracy & Randles, 2011). In contrast, the dimension of valence is part of the basis for the core-affect perspective that holds that emotions are perhaps best understood on a spectrum like valence (Russell, 2003). It is not a problem that there are such competing views on emotion. The problem is that researchers can be disconnected from a respective theory on emotion to the extent that they reinterpret results in theoretical frameworks that are opposing theoretical views from the basis of the measures they are using - which can introduce confusion into the literature (Weidman et al., 2016).

In addition to poor alignment between emotional theory and measure, at times, researchers did not explicitly state their theory on emotion (Weidman et al., 2016). This lack of declaring a position results in studies where the theoretical perspective on emotion can at best be inferred by how such researchers determine accuracy of their measures. This underscores the problem that researchers using emotional measures are not always explicitly thoughtful in terms of aligning emotional theory with studies that measure emotion.

To provide a specific example of this in this PhD thesis I focus on the measure of SA which detects the intersection between the intention of the author of the text and the reaction the communication elicits by the intended reader of the text (Balahur & Steinberger, 2009). The emphasis on SA is because this measure has shown promise in educational research (Hillaire et al., 2018; Wen et al., 2014) while simultaneously raising concerns about the accuracy and potential limitations of the SA approach (Hillaire et al., 2018; Wen et al., 2014). While some researchers interpret SA results cautiously there is evidence that some researchers adopt less nuance when interpreting results of SA.

For example, in one study (Rajput et al., 2016) where SA was used to detect if written expression in teacher evaluations was positive, negative, or mixed, the researchers did not explicitly state their theoretical perspective(s). However, they checked the accuracy of the measure against the researchers' perceptions of what is "*actually*" positive and negative text. They found the best performance occurred in

agreement on positive labels, with lower levels of accuracy for negative and mixed (positive F-Score=0.95; negative F-Score=0.67; mixed F-Score=0.57). This takes the perspective that there is a truth that can be perceived by researchers in terms of what text evokes positive or negative emotion. These results indicated that their SA classifier agreed with researchers' perceptions of emotion expression in text. The belief that researchers can identify positive and negative text aligns better with universal perspectives reflected by their state of labelling the actual valence score themselves.

At the same time, in the Rajput, Haider, & Ghani (2016) study SA was also used to compare their sentiment detection of text from teacher evaluations with the Likert score the student used to answer a series of questions about student satisfaction. By evaluating the accuracy of detecting text in teacher evaluations against the Likert score provided by students (by mapping scores to positive, mixed, and negative evaluations) they seem to also subscribe to a social perspective on emotion. When comparing Likert scores with SA on text from the evaluation they found a correlation of 0.64. As they benchmarked against a related set of questions scored using a Likert scale they fall short of asking students if the text they wrote is positive, negative, or mixed, and used the Likert scaled scores as a proxy of the student opinion. This aligns with the origins of SA, which examined movie reviews in a similar fashion by considering the text of the film review and an associated Likert score of a film (Pang & Lee, 2006). Given that the Rajput et al. (2016) study used both researchers' perceptions to evaluate accuracy of teacher evaluations and student perception to evaluate accuracy, the study does not seem to explicitly focus on an emotional theory, but rather seems to consider multiple perspectives to determine accuracy. The potential disconnect between theoretical perspective and measurement is precisely why it is important to undertake work that focuses on rigorously examining the accuracy of SA based on an explicit theory of emotion. If SA is intended to measure the opinion of the author and the reactions of the intended audience, then it may be of more interest to consider how this measure relates to what students think is positive and negative when reading the text of their evaluation.

Alternatively, some researchers have focused on how emotion expression in text during learning related to outcomes, placing more of an emphasis on how emotion expression relates to student goals. For example, one common use of SA in the context

of learning is to see if emotion expressed in discussion forums of a course relates to course completion (Chaplot, Rhim, & Kim, 2015; S. Crossley, Paquette, Dascalu, McNamara, & Baker, 2016; Wen et al., 2014). When studies focus on how emotion expression in course discussions relate to course completion the emphasis is on how emotion relates to goals, in this case completing the course, with emotion expression. However, Chaplot et al. (2015) did not explicitly state a theoretical perspective on emotion. Wen et al. (2014) outlined how positive attitudes are important for success in e-learning contexts. Wen et al. (2014) also stated that there are some specific negative emotions, such as boredom and frustration, that are prohibitive to learning. Finally, they articulated that emotion expression must be interpreted with nuance as highly engaged students can still make negative comments (Wen et al., 2014). While this nuanced perspective on interpreting SA results demonstrates a thoughtful perspective based on prior evidence, it fails to explicitly state an underlying theory of emotion and how that theory relates to emotion expression.

While aligning emotional theory and measurement is important, another potential barrier towards investigating the role of emotion in learning using contemporary measures such as SA is that there is a close relationship between how a learning environment is designed and what information is available to measure. Indeed, some have outlined that learning analytics in large part relies on ‘exhaust’ which is a by-product of learners interacting with their environment as investigating data which is largely comprised of a by-product of learners interacting with each other and learning activities (Shum & Crick, 2012). The handbook of learning analytics acknowledges this (Lang, Siemens, Wise, & Gasevic, 2017, pg. 129) as a streetlight effect (Freedman, 2010) where investigations can focus on what data is available as a convenience over focusing on data that is ideal. Effectively if I do not design a light to shine on emotional dimensions of learning I may be investigating the role of emotion using exhaust (e.g., lights shining on the wrong things) rather than directly investigating the interplay between emotion and cognition during learning. This perspective supports the need to consider the relationship between learning design and learning analytics (Rienties et al., 2017).

1.3 PROPOSED SOLUTION

An important aim of this thesis is to establish a robust SA measure that stays close to the theoretical perspective that emotions are socially constructed (Barrett, 2012) in a collaborative learning context where learning can also be examined as a social construction (Kalpana, 2014; Van Den Bossche et al., 2006; Vygotsky, 1978). In addition to theoretical alignment between emotion and learning I also aim to align design by attempting to integrate the sub field of learning design (Bakharia et al., 2016) - emotional design (Uzun & Zahide, 2018). To put it succinctly, research on emotion in learning is filtered through the measures of emotion influenced by competing theories (Green et al., 1993) which are intertwined with the design of the learning context (Rienties et al., 2017). A key set of challenges to investigating the role of emotion in learning using analytics focused on emotion include: articulating a theory of emotion and proposing a measure that aligns with that theory (Weidman et al., 2016), proposing an alignment between the learning goal and learning design (Wise & Vytasek, 2017). When considering both of these key challenges in concert there is a need to coordinate emotional theory, learning theory, and learning design when investigating the validity and reliability of an emotional measure in the context of learning.

This thesis has selected the context of online group work mediated through online text communication for a multitude of reasons. First, there is a link between social and emotional aspects of learning (Immordino-Yang & Damasio, 2007; Ludvigsen, 2016; Schachter & Singer, 1962), and a group work context provides an opportunity to examine the role of emotion during collaboration through participation in online text based conversations. Collaborative learning also offers social theories on learning, including self and socially regulated learning, which considers how people learn in social contexts providing a learning theory that acknowledges the role of social interaction in learning. I investigate the role of emotion in learning in this context using new affordances of measurement (Rienties & Rivers, 2014). Specifically, I focus on the emerging work in emotional learning analytics (ELA) where researchers are exploring how emotion expressed in text communication can provide insight in the role of emotion in learning (Lang et al., 2017).

There have been many limitations and challenges uncovered in existing research on SA in learning as many studies have cautioned that there are challenges when interpreting SA results due to concerns over accuracy (Altrabsheh, Gaber, & Cocea, 2013; Calvo & Kim, 2010; Chaplot et al., 2015; Kagklis, Karatrantou, Tantoula, Panagiotakopoulos, & Verykios, 2015; Koehler, Greenhalgh, & Zellner, 2015; Munezero, Mozgovoy, Montero, & Sutinen, 2013; Santos, Salmeron-Majadas, & Boticario, 2013; Wyner, Shaw, Kim, Li, & Kim, 2008). Furthermore, there is some suggestive evidence that emotion expressed in text in online discussions can correlate with learning outcomes in situations where students are outperforming peers (Hillaire et al., 2018). However, some of these studies (Hillaire et al., 2018; Wen et al., 2014) take a cautious approach toward interpreting results cautioning that the accuracy of SA requires a more thorough investigation. By closely examining previous methods for SA in the context of learning and considering the kinds of data analyzed, I aim to generate a SA measure which can help understand the role of emotion in learning.

Given the multitude of studies which caution that SA technologies generally lack accuracy in the context of the learning, I emphasize an exploration of validity of a SA measure in a context of two large cohorts of 800+ business students to establish a solid foundation upon which to explore emotion in learning using SA. The approach towards establishing a valid measure in the context of learning is to train a machine learning classifier based on the perspective of students effectively aligning the student perspective and the accuracy of sentiment analysis in social terms. In other words, by using the student perspective to determine what is “actually” positive and negative I refer to this as a student sourced sentiment analysis classifier (SSSAC).

As the labeled data in this case comes from a crowd there is reason to question the reliability of rating. Typically, crowdsourcing methods yield lower levels of agreement so it is suggested to examine inter-rater reliability through reporting both Fleiss’ Kappa and Krippendorff’s alpha when using crowdsourcing methods. Some have suggested that crowdsourcing should stop considering the crowd as anonymous and detached and instead consider how the crowd selected influences their ratings. In this thesis is the crowd is not anonymous and detached from the context, but instead the crowd whose opinions I are attempting to model – the students. This complicates the analysis of

agreement statistics further as low levels of agreement from crowds might simply suggest valid disagreement of the crowd.

As the aim of SA is to model the opinion of the author and reaction elicited by the intended audience the opinions I collect from the students are the aim of the measure. As students provide examples in a method that builds on crowdsourcing approaches it is important to examine inter-rater reliability to evaluate the quality of the examples. In conjunction with examining the extent to which students agree I also examine the stability of student opinions through test-retest reliability. Understanding the stability of opinion through examining the extent to which students provide the same labels to text at two different points in time I aim to better understand the extent to which student opinions are stable. Finally, as the crowdsourcing of student opinions is not common practice I compare this approach to ground truth by using Mechanical Turk to crowdsource labels for the same messages and compare the extent to which MTurk ratings agree with student ratings.

As discussed in the three models of emotion earlier if SA is connected to the BTE then valence in text would be identifiable by researchers from outside of the classroom context where as if SA aligns more with the CTE then only members of the classroom context would be a part of the social context and have the necessary insider perspective to select an appropriate label for text messages generated in the classroom. This raises research question 1 (RQ1): To what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters? RQ1A: To what extent do students agree in terms of inter-rater agreement when providing examples? RQ1B: To what extent do Mechanical Turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples?

While RQ1 examines student generated ground truth with researcher generated ground truth the inevitable differences of opinion do not themselves prove one is superior to another, but rather examine the extent to which differences arise. To understand the comparative utility of researcher and student opinions I next train a classifier based on both of those opinions and see which classifier does a better job at predicting a novel set of student opinions from a new cohort of students. To setup such a comparison I first must establish that the generation of a domain specific classifier was

effective. The literature suggests domain specific classifiers should outperform general classifiers so I first examine if a classifier generated using student opinions does a better job than general classifiers to examine if the minimum threshold of validating this domain specific classifier was effective.

When training a new machine learning classifier one of the established ways to measure accuracy is to take the labelled data (Browne, 2000; Little et al., 2017), in this case the student sourced labels, and train the classifier using a training data comprised of 90% of the sample and seeing how accurate the classifier is on the test data which is comprised of remaining 10% of the data. Then repeating this process ten times so that every message is in the in the test subset of data once. This process is referred to as 10-fold cross validation. While cross validation provides a prediction of how accurate a classifier will be when interacting with novel data, it is important to confirm that prediction by collecting a new set of data and using the classifier that has been cross validated to see if it achieves the expected level of accuracy on new data. A common way to evaluate the accuracy of a cross validated classifier is to replicate the process used to generate the labelled data (Browne, 2000; Little et al., 2017). By conducting a study and replication of that study I explore research question 2 (RQ2): To what extent can crowd sourced, and in particular student sourced, examples train a machine learning classifier to predict the valence categories of positive, negative, neutral, and mixed? RQ2A: To what extent can student labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2B: To what extent can Mechanical Turk labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2C: How do logistics classifiers trained using student labels and Mechanical Turk labels compare to general benchmarks when predicting the valence categories of positive, negative, neutral, and mixed? RQ2D: To what extent do students find predictions from a student sourced classifier useful?

To collect data to answer RQ2 it was necessary to conduct a replication study. When conducting replications, it is important to consider how a systematic replication can help to confirm results. For example, using an experimental condition intended to improve results in combination with a control condition the replication can both confirm expectations, using the control condition, and provide evidence of validity, with an

intervention, if the expected improvement occurs. To consider how to construct an intervention that would improve the student detection of emotion expression I turn to the literature on design because of the relationship between design and analytics (Rienties et al., 2017; Wise & Vytasek, 2017).

For this thesis I have adopted the Universal Design for Learning perspective as it puts emotion as a central aspect to learning (Rose, 2015). Part of the rationale that this is a useful exercise is that contemporary neuroscience suggests the interplay of emotion and cognition is likely important to complex processes such as learning (Immordino-Yang & Damasio, 2007; Okon-Singer, Hendler, Pessoa, & Shackman, 2015). To clarify that this thesis is not neuroscience research I leverage the Universal Design for Learning (UDL) framework to translate what is understood about neuroscience to inform how I design learning environments (Coy et al., 2014; Dinmore & Stokes, 2015) to generate learning analytics from trace data in UDL environments (Hillaire, Rappolt-Schlichtmann, & Stahl, 2014). One of the strategies adopted by UDL when supporting online text communication is sentence starters. This raises research question 3 (RQ3): To what extent can emotional sentence starters improve the inter-rater reliability of student examples?

While training a classifier using cross-validation to predict the accuracy and following up with a new data set to confirm expected accuracy provides insights into the internal reliability of the classifier, it is also important to understand the extent to which the method is reproducible. Collecting new labelled data can be used to examine predicted accuracy. It can also be used as a new training set of data to cross-validate the measure using the newly collected data to examine if the process of training the classifier is replicable. Demonstrating the reproducibility of the SSSAQ approach raises research question 4 (RQ4): To what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?

While RQ1-RQ4 focus on providing insights into the ability to predict student perceptions of emotional communication in text, the next step I explore is how such a measure relates to existing emotional measures. I explore additional measures such as React – A self-report of emotional response (Hillaire et al., 2018) administered right

after the warm up exercise, as well as traits of emotions, such as commonly measured by emotion instruments like the Berkley Expressivity Questionnaire (BEQ) of Gross and John (1997) administered ahead of the lab and the Positive and Negative Affect Schedule (PANAS) instrument developed by Leue and Beauducel (2011) administered after the lab. . This is an important question, since if SA does not measure the same underlying phenonema of emotion addressed by psychometric measures, it raises important questions about what underlying phenomenon is being captured by SA. This raises research question 5 (RQ5): To what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?

To provide a baseline comparison of how a general SA technology relates to the aforementioned emotional measures I examine the SA technology called Sentistrength (Thelwall, 2013; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) that did not use student sourcing to generate labelled data. The reason to compare another SA technology is to frame the interpretation of the measurement by comparing to a SA that was built for general purposes. I selected SentiStrength as the comparison technology because it was created using the same theoretical basis of emotion. SentiStrength considers positive and negative emotion to be both parallel experiences, where people can feel positive and negative at the same time, as well as integrative experiences, where emotion can at times integrate positive and negative aspects, resulting in a polarized experience that is either positive or negative (Cacioppo et al., 1999). By examining the extent to which a comparable technology has correlations with emotional measures this raises research question 6 (RQ6): To what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?

In summary, the main aim of this thesis is to investigate training a SA classifier using labels generated by students and examining the accuracy of the measure using cross-validation and replication. By conducting a systematic replication with a control and intervention condition I test if an intervention, designed to support emotional communication, has the intended effect of increasing the accuracy of the measure to help determine if the results of accuracy represent the intended construct - student

perception of emotional communication in text. Afterwards, I examine the external validity by comparing SA with psychological measures of emotion to determine the extent to which SA trained on student perception provides a measure that aligns with the emotions of students. In other words, the purpose of this PhD thesis is to take a critical perspective on the validity of SA in ELA by considering internal reliability (using cross-validation and replication) and external validation (using psychological measures) to determine the extent to which SSSAC can provide insight into the role of emotion in learning. To conduct this investigation, I have selected a social theory on emotion, a social theory on learning, and a learning design approach that emphasizes the role of emotion in learning. While this selection frames the investigation as positioned heavily towards social perspectives it is deemed necessary to explicitly state the theoretical underpinnings of the investigation to appropriately articulate the alignment between emotional theory and measurement as well as emotional measurement and learning design.

1.4 THESIS STRUCTURE

1.4.1 CHAPTERS

1.4.1.1 CHAPTER 2 LITERATURE REVIEW

In Chapter 2, I have organized literature into two main sections: 2.1 Theory and Measurement of Emotion; 2.2 Theory and Design of Learning. As the focus of this thesis is on creating and evaluating a measure of emotion for the context of learning there is more emphasis on section 2.1 than there is on section 2.2. In section 2.1 I review three models of emotion, the components of emotion, and a theoretical perspective on emotion expression. I then detail measures of emotion expression by considering the theory on categorizing emotions as positive and/or negative by discussing multiple approaches towards measure valence. I propose using the Univariate Mixed Emotion (UME) model which considers emotion to be best categorized as positive, negative, neutral, and mixed (both positive and negative). With a refined model of emotion, I propose a measure that is created based on this theoretical perspective in the context of learning using student perception as the foundation for accuracy of the measure. Section

2.2 outlines design and theory of learning for the expressed purpose of considering how to design a learning context to support emotion expression. The purpose for designing learning supports for emotion expression is to support the validation of the proposed measure detailed in Chapter 3 Methodology.

1.4.1.2 CHAPTER 3 METHODOLOGY

In Chapter 3 I continue with these elements to consider affect, feeling, and emotion from the ontological perspectives of objectivism, subjectivism, and social constructivism. By focusing this work on social aspects of emotion, I connect with the social constructivism theory on learning. These theoretical perspectives generate implications for determining that accuracy of a SA measure based on the UME model. In considering how to evaluate this model a pragmatic epistemic perspective is adopted. Furthermore, Design-Based Research methodological approaches (DBR) are used to investigate the accuracy of a cross-validated measure. In this section I examine UDL guidelines for design and outline design proposals that consider the interplay of emotion and cognition in support of the development and evaluation of an emotional measure for the context of learning. To examine external validity of the measure I consider psychological measures that consider emotional states and traits of students.

1.4.1.3 CHAPTER 4 STUDY 1

In Study 1, I examine two research questions: RQ1: To what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters? RQ1A: To what extent do students agree in terms of inter-rater agreement when providing examples? RQ1B: To what extent do Mechanical Turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples? This compares agreement of students to a standard crowd sourcing approach. I also examine RQ2: To what extent can crowd sourced, and in particular student sourced, examples train a machine learning classifier to predict the valence categories of positive, negative, neutral, and mixed? RQ2A: To what extent can student labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2B: To what extent can Mechanical Turk labels train a logistic classifier

which predicts the valence categories of positive, negative, neutral, and mixed? RQ2C: How do logistics classifiers trained using student labels and Mechanical Turk labels compare to general benchmarks when predicting the valence categories of positive, negative, neutral, and mixed? RQ2D: To what extent do students find predictions from a student sourced classifier useful? In doing this comparison we examine if students do a better job than Mechanical Turk in generating labels to train a classifier and benchmark against general approaches as specialized classifiers should do better than general methods.

To explore the approach of student sourcing a classifier Study 1 uses data from two groups of students. The first group is from 2016 and had 767 students participating in an online group collaborative activity and then reflect on the emotional aspects of the group discussion. Students review their own group discussions and provide examples of messages for each of the valence categories of positive, negative, neutral, mixed, and ambiguous. The crowd sourcing examples result in a set of messages with what is believed to be the label that best reflects the student experience of the messages as judged by member of the social context where the messages were authored and read. The second group is from 2017 and had 447 students replicate the activity from 2016.

The results of RQ1 indicate that the student-sourced labeling demonstrates moderate reliability which is better than Mechanical Turk with fair agreement and I interpret student labels sufficient for educational research. The results of RQ2 demonstrates that training a classifier based on student labels is better at predicting future student labels than a classifier trained on Mechanical Turk label as well as general benchmarks. While the overall accuracy of the student sourced classifier is lower than many published studies on sentiment analysis the significant contribution is that this measure is evaluated on student opinions which appear to be diverse as indicated by RQ1. I argue that the challenges in accuracy this approach faced are challenges any SA technology should reconcile as evident by the lower performing general benchmarks which are used in published studies in educational research.

1.4.1.4 CHAPTER 5 STUDY 2

In Study 2, I turn my focus on improving the process of student sourcing through the use of the emotion awareness tool of emotional sentence starters (ESS). I examine the 2017 group of 447 students as a control condition (2017C) and introduce data from a second 2017 group of 437 students in an experimental condition (2017SS). The students in the control and experimental condition are all from the same class and were randomly assigned.

In the experimental condition, students are asked to express their reaction to the data in the activity using ESS at least twice during the discussion. Students were asked to use one of four emotion sentence starters: “I had a positive | negative | neutral | mixed reaction to...” at least twice during their group discussion.

When answering RQ3: “To what extent can emotional sentence starters improve the inter-rater reliability of student examples?”, I first examine participants in the sentence starter condition and remove all examples which used the ESS leaving solely unscripted examples. Then I compute inter-rater reliability and compare that the inter-rater reliability from the control condition. When examining the subset of examples from the emotional sentence starter condition, excluding the examples where the supports are used, I find dramatic increase in inter-rater reliability result from the intervention of emotional sentence starters.

When answering RQ4: “To what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?”, I find when using examples from the emotional sentence starter condition, excluding the scripted examples, I find dramatic improvements in terms of inter-rater reliability achieving substantial agreement as compared with moderate agreement in the control condition. I also find that comparing the same number of student examples from the 2016 data with data from the 2017SS data restricted to unscripted examples improved the accuracy of a classifier, I also find evidence that more data would likely improve the classifier further by combining data from the 2016 examples with the 2017SS unscripted examples.

This experiment contributes the use of ESS to improve student sourcing an SA classifier. These results also raised questions about what additional effects may result when student awareness about the sentiment of group communication increases.

1.4.1.5 CHAPTER 6 STUDY 3

In Study 3, I examine the external validity of a SSSAC using (n=868) students who both participated in either the control or experimental condition from Study 2 and provided responses for all of the psychological measures used for correlation analysis. When answering RQ5: “To what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?”, I compare predictions of emotional expression from a SSSAC with a wide spectrum of psychological measure of emotional state and trait. When answering RQ6: “To what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?”, I conduct a parallel investigation into the external validity of a SSSAC that was not trained on student sourced examples as a benchmark. The results indicate that neither SSSAC or SentiStrength sentiment analysis methods correlate with the emotional measures examined.

1.4.1.6 CHAPTER 7 CONCLUSIONS

In the conclusion Chapter 7 I will discuss the results of our analysis of a SSSAC in terms of accuracy for which I find compelling evidence, and correlation analysis for which I find limited evidence, to critically consider if and how SA might be used in the context of learning. I contrast the accuracy results with the benchmark technologies to explore the extent to which the student sourced classifier is valid in terms of student perception and the extent to which alternative approaches align with student perception in the context of learning. These results provide insights into the strengths and limitation of SA to provide a measure which aligns with student perceptions for the valence categories of positive, negative, neutral, and mixed. I consider the extent to which the design of learning technology influences the measurement of emotion in terms of accuracy based on student perceptions.

Based on our results I discuss the importance of learning design in terms of the accuracy of emotional measurement focused on predicting student perceptions of emotion in text communication. Then I contrast the results of limited evidence of

correlation with other emotional measures by discussing the extent to which a comparable SA technology, SentiStrength, demonstrates a similar lack of correlation. By considering the results I connect back to the theory on emotion and emotional measurement to make sense of the findings.

Overall, my main scientific contribution to the field is that student perceptions can be used to create a SA classifier to predict student perceptions of valence categories of positive, negative, neutral, and mixed expressions better than Mechanical Turk perceptions examined in this thesis. Along with evidence that it is possible to train a SA based on labels provided by student perceptions I also uncover that student perceptions of emotion expression in text does not appear to have a relationship with the emotional experience of the author of the text. This evidence suggests that at least in circumstances similar to the experiments in this study (one-hour lab activities) emotional expression appears to be independent of emotional experience.

The main take away from the results of this PhD thesis suggests that when students look at text expression and consider emotional expression this does not necessarily provide an effective line of sight on the emotional experiences of their peers. Within this main takeaway there is a suggestion that while cross-validation and replication can provide clear results for the accuracy of a SA classifier, further future research needs to be conducted in terms of exploring how SA relates to the emotional experience of students to determine exactly which facets of emotion SA are measuring. In this PhD thesis I find that the SSSAC is not measuring the internal emotional experience of the author. Future work should examine to what extent SA measures a social facet of emotion.

CHAPTER 2 LITERATURE REVIEW

In this Chapter 2 I will first review three broad concepts of emotion in section 2.1, and how these concepts can be used to measure emotion. In particular, I will review how emotions could potentially be measured by reviewing recent sentiment analysis (SA) literature. Afterwards, in section 2.2 I will review how learning scientists develop potential intervention strategies to effectively support learners in online environments.

2.1 THEORY AND MEASUREMENT OF EMOTION

In this section I highlight three theoretical perspectives on emotion. The reason that three frameworks are discussed is to contrast a few different points of view from a theoretical perspective of emotion so that they can be referenced when reviewing previous research on SA in the context of learning. As indicated in Chapter 1, several researchers use emotional measures that may not always explicitly state their theoretical position on emotion (Weidman et al., 2016). Therefore, I review three contrasting models of emotion as a foundation for interpreting existing work. The three models I review are: Basic Emotion Theory (BET) which defines emotion as clearly defined relationships between physiological response, action, expression, and experience; emotion as a Constructed Theory of Emotion (CTE) which defines emotion in terms of a social consensus; and SAT which defines emotion as closely tied to goal achievement. Each model is described and the limitations of the model are briefly discussed. After defining these three models, I categorize previous work on SA in the context of learning by selecting the theoretical model(s) of emotion which best fit how researchers have previously evaluated accuracy of SA.

Basic emotion considers some emotional experiences to be so fundamental that they are described as universal. For example, people may have a common experience of emotion when it comes to some specific emotional responses, such as anger and happiness. Typically, researchers who adopt the BET perspective focus on 5-7 emotions that are considered fundamental to the human experience. In a review of four models of basic emotions the models suggest that emotions that might be so fundamental that they

are universal include: Happiness, Enjoyment, Sadness, Fear, Anger, Disgust, Interest, Contempt, Rage, Love, Lust, Care, Surprise (Tracy & Randles, 2011). Basic emotions start with a stimulus that evokes an emotional response, which is comprised of non-verbal expression, physiological response, and behavioral preparedness (Tracy & Randles, 2011). I can unpack the components of basic emotion using anger as an example. Anger starts with a stimulus that triggers anger. For example, if a student read a disparaging comment about themselves from a peer this may trigger anger. This anger response may be visible in their facial expression, their heart-rate and breathing might change, and they might be preparing to take action like retaliation for the disparaging remark by replying with a remark that escalates the conflict.

One limitation for basic emotion research is that there is minimal relevance for basic emotions in learning activities that span 30 minutes to 2 hours (Calvo & D'Mello, 2010), which is the time-span for the experiments designed in this PhD in Chapters 4-6. For example, if I examine the anger example of the online comment I could interpret the minimal impact this has in the context of a one-hour learning activity by considering that this type of stimulus and response may not be prevalent in a one-hour learning activity. The basic emotions (e.g., Happiness, Fear, Anger, Disgust) may not be prevalent in a one-hour learning activity. In summary, BET lends itself to measurement through strongly defined relationships between emotion and behavior, but the basic emotions are not thought to be particularly relevant to learning during 30-minute to 2-hour learning activities.

Constructed Theory of Emotion (CTE) is a perspective that suggests that the manner by which emotion is interpreted is through the influence of social factors. An example of how social theorists interpret emotion is illustrated in the book "How Emotions Are Made" by Lisa Feldman Barrett when she uses a picture of Serena Williams. The photo was taken immediately after Serena beat her sister, Venus Williams, in the 2008 U.S. Open. The picture Barrett presents is a cropped image of Serena's facial expression and Barrett suggest that looking at the facial expression in isolation of context might be categorized as an expression of terror when using a basic perspective on emotion. However, by taking context into consideration I should instead interpret the image to mean something closer to exultation (Feldman Barrett, 2017, pg.

42). The full image depicts Serena making a fist in the air as part of her physical response to the victory. Barrett argues that emotion is comprised of making meaning, prescribing action, regulating the body, emotion communication, and social influence. The three components of making meaning, prescribing action, and regulating the body are considered individual as they can all be done in isolation. Two of the components, emotion communication, and social influence are considered social as they are aspects of emotion that cannot be done in isolation as they are social interactions. To return to the example of Serena Williams as she wins the U.S. Open she makes meaning of the win, and regulates her body resulting in a facial expression and the physical action of making a fist in the air. These physical cues demonstrate social communication of a celebration of victory to the audience of the U.S. Open. It is through the context of the U.S. Open that the audience interprets the communication as a positive emotion expression. One limitation toward taking a social reality perspective on emotion is that it requires considering the perspective of multiple people familiar with the social context to determine emotion. Whether this is in practice feasible in our PhD context (see Chapter 3) can be debated.

Situated Affectivity Theory (SAT) considers the goal as the focal point for interpreting all of the components of emotion (Wilutzky, 2015). With this goal orientation a manipulation between an individual and their environment is the basis for a stimulation for emotion. The physiological response represents a physical experience that resonates with the interaction with the environment. Emotional communication is thought to be used by people to achieve goals. Actions are thought to be intertwined with emotion as they are both closely linked with goal orientation.

To summarize these three perspectives on emotion discussed in this review I have highlighted how some researchers:

- 1) consider emotions as predictable relationships (BET)
- 2) consider emotions as social constructs (CTE)
- 3) consider emotions to be goal oriented (SAT)

These three perspectives are highlighted in Figure 2.1 (with an associated quote that represents these theoretical perspectives). While these three theoretical models are not

exhaustive they are divergent enough that I can identify how previous work aligns with them to help organize previous work in SA (see section 2.1.3.1)

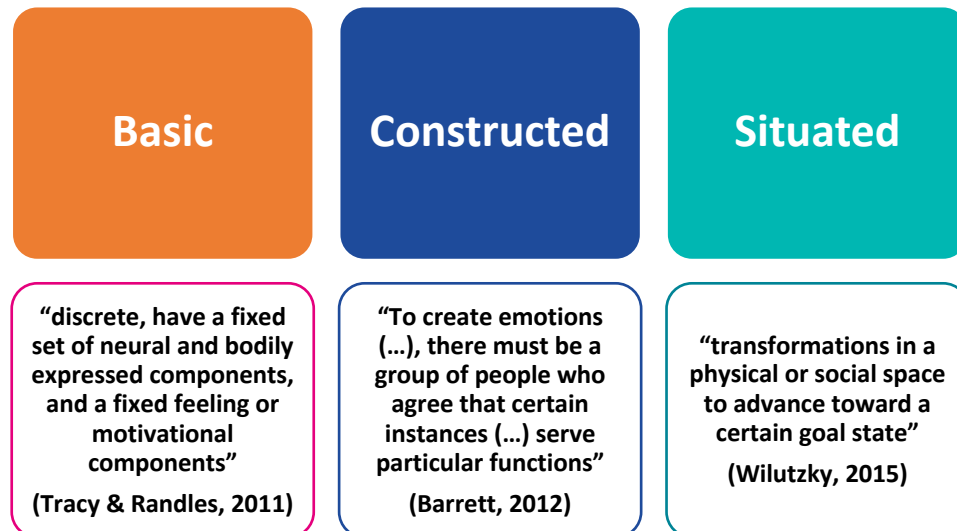


Figure 2.1 Three theoretical perspectives on emotion: Basic, Constructed, and Situated

While three models above have different perspectives on emotions there are some elements of agreement between them. In the next section I detail the multi-component model of emotion and illustrate how BET, CTE, and SAT map into a common set of components of emotion to contrast the theories. The purpose for contrasting these three theories is to establish a focus in this thesis on the theoretical perspective that consider emotion as a social reality.

2.1.1 COMPONENT PROCESS MODEL OF EMOTION AND COMMUNICATION

While there are many disagreements among emotion researchers to frame the three emotion models above I compare all three models with the Component Process Model (CPM) of emotion, as there is a theoretical discussion of the intersection between emotional communication and components of emotion (Scherer, 2009). The CPM model has been developed over 25 years and has many studies that support the proposed structure (Scherer, 2009). As indicated in Figure 2.1.1., the CPM model of emotion distinguishes between Conscious representation and regulation (C), Unconscious

reflection and regulation (U), and Verbalization and communication of Emotional Experience (V). The CPM details how the majority of psychological studies focus on the verbal account of consciously experienced feeling (which occurs at the intersection of circles C and V in Figure 2.1.1). There are three types of emotional communication in this diagram that I consider. Scherer labels the intersection of communication with the conscious representation and regulation and unconscious reflection and regulation as Valid Self-Report of Emotion (X1 in Figure 2.1.1). I expand on this to label the intersection of communication of emotion and conscious representation and regulation as Regulatory Expression (X2 in Figure 2.1.1), and I label the communication of emotion that does not intersect with conscious or unconscious regulation as Disconnected Expression (X3 in Figure 2.1.1).

When asking about emotional experience through self-report the responses are considered to be valid (i.e., Valid Self-Report of Emotion) when they occur at the intersection of verbalization and communication of emotional experience, with both the conscious representation and regulation and the unconscious reflection and regulation. The reason this is considered valid from the CPM perspective is that all of the components of emotion that occur at the unconscious reflection and regulation are frequently referenced components across emotional theory. The components detailed in the CPM are physiological symptoms, cognitive appraisal, motor expression, and action tendencies. When there is alignment between Conscious representation and regulation (C) with these components of Unconscious reflection and regulation (U), which is expressed through Verbalization and communication (V) this is considered emotional expressions that accurately reflect the emotional experience of the person.

As I know there are limitations with self-report of emotion, as people may provide responses for a whole host of reasons that may not reflect their internal state (e.g., providing socially desirable answers) (Duckworth & Yeager, 2015). Another possible breakdown is that emotional communication may simply be intended to misdirect the audience, which could indicate Disconnected Expression. However, if a speaker communicates a socially desirable response as part of conscious regulation wherein the speaker is trying to convince themselves to have an expected emotional experience, this

would be categorized as Regulatory Expression. Considering the intersection of unconscious reflection and regulation with communication is out of scope for this thesis.

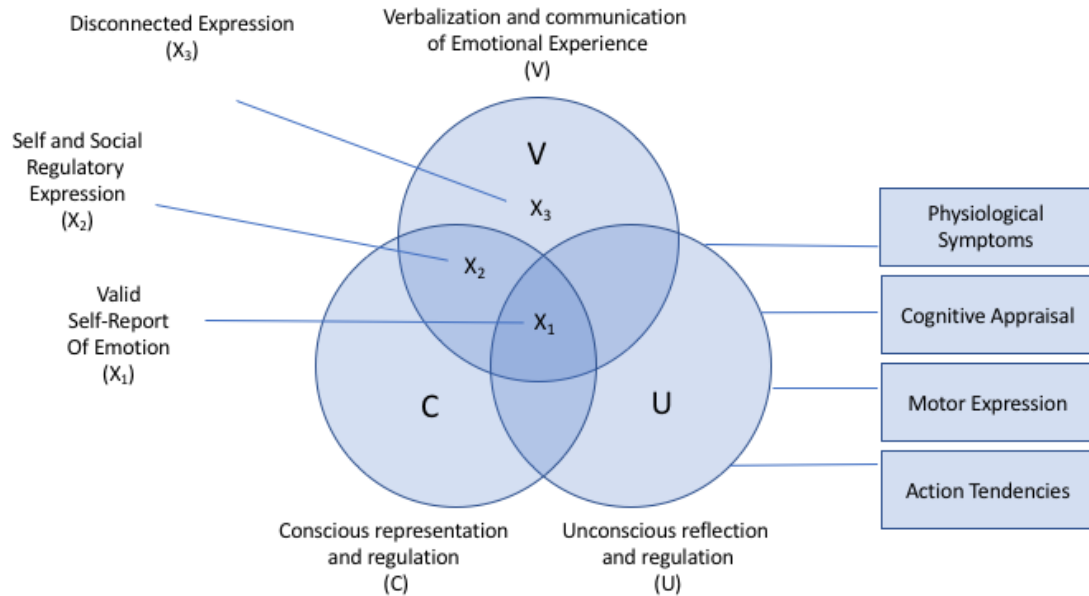


Figure 2.1.1 Component Process Model (CPM) of emotion adapted from Scherer (2009)

The focus of this thesis is exploring the extent to which emotional communication intersects with Valid Self-Report of Emotion. To investigate this intersection, I consider how validated measures of emotion have correlates with measures of emotional expression detailed in section 3.2.3.1. As the CPM does not label or detail Regulatory Expression or Disconnected Expression (labels I have added to the diagram) I categorize emotional communication detailed in the CTE, BET, and SAT as a means to provide a theoretical position for Regulatory Expression as well as Disconnected Expression.

In the CTE one form of expression is emotional communication where I synchronize our perspectives using communication (Feldman Barrett, 2017, pg. 139). For example, if Person A were communicating their emotional experience, then Person B might ask clarifying questions to help synchronize their interpretation of the communication. In asking clarifying questions Person B would be participating in a self-regulatory process to align their emotional perception with the intention of Person A. The communication from Person A would be classified as Valid Self-Report of Emotion once Person A and

Person B achieved synchronization using Regulatory Expression. The other form of emotional communication detailed in the social theory of emotion is that of Social influence (Feldman Barrett, 2017, pg. 139). When engaging in communication of social influence Person A would be explicitly trying to influence Person B with their communication. Where social influence diverges from Valid Self-Report of Emotion is that the requirement of social influence is that the communication must be part of a collective intentionality. Collective intentionality suggests that no matter how meaningful an expression is for Person A, unless it is part of a collective perspective on emotion it perceived as just meaningless noise (Feldman Barrett, 2017, pg. 139).

So as two individuals navigate their communication toward a collective intentionality, it is reasonable to assume that one form of conscious representation that is not necessarily a valid self-report from the perspective of Person A's may be used to achieve a collective perspective on emotion to achieve a Valid Self-Report of Emotion. In terms of the CTE a disconnected expression might be best described as communication that is not intended to be emotional that is perceived as emotional communication. For example, Barrett provides an example where a man is stamping his boots as he walks because he is trying to get dirt clods off of his shoes. In this example, an observer perceives this as a communication of anger. From the CTE it would not be considered anger until there was a consensus reached (Feldman Barrett, 2017, pg. 139). Effectively, from the CTE what makes for a Valid Self-Report of Emotion is the achievement of consensus in a social context.

In terms of SAT emotion expression is considered the same as the emotion as it is through emotional expression that an individual interacts with their social environment (Wilutzky, 2015). Given the tightly coupled nature between emotion and expression from the SAT perspective, the only emotional expression is Valid Self-Report of Emotion as indicated by the perspective that expression is "part and parcel" with emotion (Wilutzky, 2015). As the identifiable mark of emotion in SAT is the alignment between action and emotion to achieving a goal (Wilutzky, 2015) emotional expression simply represents how an individual seeks goal achievement and it is examining their actions and expressions that I can determine their emotional state.

In terms of BET, non-verbal expression is one of the fundamental criteria as basic emotions are thought to be universal, and exhibited through facial expressions that can be interpreted across cultures (Tracy & Randles, 2011). Advances in BET (Keltner, Sauter, Tracy, & Cowen, 2019) have explored multi-modal expressions including head tilts, gaze, and non-word utterances (e.g., sighs, ooh, and ahhs). However, there is not a clearly articulated notion of what people are predicted to say. This is likely because expressions with words are culturally bound and not universal.

In summary, by taking the CPM and labelling three areas of interest of emotional communication of Valid Self-Report of Emotion, Regulatory Expression, and Disconnected Expression I find that examples provided by the Constructed Theory could be interpreted for each of the three areas of interest. SAT by definition considers emotional expression to only be Valid Self-Report of Emotion. BET only connected with verbal expression in terms of non-word utterances. As the aim of this thesis is the measurement of emotional expression in text communication the CTE appears to be the best aligned with the investigation. While BET does not align with the focus of this thesis as this theory holds the notion that emotions are fixed identifiable relationships that are universal in nature (Tracy & Randles, 2011) are reflected in practices of researchers who use SA. In addition, the SAT theory which has an emphasis on goal achievement is also reflected in practices of researchers who use SA. There are many other theories of emotion, but these three have been selected as they demonstrate a usefulness when interpreting how previous researchers have measured accuracy of SA in the context of learning (see section 2.1.2.1). By contrasting these three theories to explore the suggested relationship between emotion and communication proposed by the CPM I further refined our definition of emotion. While this section has reviewed three theoretical models of emotion as a means to define three types of emotional communication I made a distinction between Valid Self-Report of Emotion and Regulated Communication.

2.1.2 WHAT IS SENTIMENT ANALYSIS?

Sentiment Analysis (SA) is commonly defined as the detection of how the opinion of the author of the text elicits a reaction from the intended reader of the text (Balahur & Steinberger, 2009). The origins of the method comes work on problems such as interpreting product and movie reviews (Pang & Lee, 2006) where a text review is written and provided in conjunction with a quantitative rating (e.g., a rating between 1 to 5 stars for a movie review). SA is a common and established approach towards detecting emotion in text expression (Pang & Lee, 2006). For example, in a review of affective computing (AC), which is a branch of computer science that aims to recognize and respond to emotional states, Calvo and D'Mello (2010) described SA as usually representing words in multi-dimensional space (MDS) to categorize text into dimensions of emotion (e.g., the dimensions of valence).

A subcomponent of AC is Affective learning (AL), which investigates how emotions affect learning based on the perspective that some affective states facilitate different kinds of thinking than others, and different kinds of thinking have long been important to research on learning (Picard et al., 2004). There is already a lot of promising evidence in AL that SA can help explore how emotion expression in text relates to learning (Lang et al., 2017; Rienties & Rivers, 2014). As part of this Chapter 2, I conducted a literature review of studies using SA in AL contexts, whereby based upon using Google Scholar, ERIC, and Web of Science, I found in total 15 studies. When reviewing these 15 studies there were many promising results. For example, some research found correlations between SA in online courses and student retention (Chaplot et al., 2015; S. Crossley, Paquette, et al., 2016; Wen et al., 2014). Several studies also indicated that SA could be used in conjunction with self-report to gain insights into the student experience while learning (Calvo & Kim, 2010; Rajput et al., 2016; Santos et al., 2013). Furthermore, researchers are starting to explore how SA can highlight for students the emotion expressed in online chat (Ortigosa et al., 2014). In order to better frame the various SA in AL approaches, the next sections will review the various mapping of emotional theories using in SA.

2.1.2.1 MAPPING EMOTIONAL THEORY TO SENTIMENT ANALYSIS STUDIES IN THE CONTEXT OF LEARNING

SA research shows promise regarding investigations into the complex role of emotion in learning. Given the potential for SA in educational research, it is essential to consider the validity and reliability of SA. To begin considering validity and reliability it is essential to precisely clarify what SA purports to measure. As it is common for researchers to use emotional measures without explicitly stating their theoretical perspective on emotion (Weidman et al., 2016), first I reviewed the 15 identified SA in AL studies in the context of learning. As indicated in Table 2. 1.2.1., the 15 studies used a range of technologies to capture text, including discussion forums, Facebook posts, diaries, self-reports, and evaluation.

Subsequently, I classified five SA in AL studies as using methods that are best described as BET, when the researchers believed that they could identify what was accurate as this indicated that emotion expression was identifiable by someone other than students in the context of learning. For example, studies in this category included an examination of teacher evaluations where researchers read the teacher evaluations, and coded the ‘actual’ sentiments based on the perspective of the researcher reporting an overall accuracy of 86.28% (Rajput et al., 2016). In another study researchers coded messages of students they reported 95.63% accuracy for positive messages; 83.51% for neutral messages; and 79.33% for negative messages (Ortigosa et al., 2014). When researchers manually coded student messages as positive or negative they reported accuracy of three classifiers and their best classifier had a precision of 0.75; recall of 0.73; and F-Score of 0.74 (Troussas, Virvou, Espinosa, Llaguno, & Caro, 2013). When comparing accuracy of student messages of 13 human raters compared with SA produced an overall precision of 0.65; recall of 0.45; and F-Measure of 0.44 (Hillaire et al., 2018). When comparing student messages with researcher labels there was a weak correlation of 0.3 reported (Santos et al., 2013).

I classified five studies all using discussion forums as reflecting Situated Affectivity when the focus was on correlations between SA and outcomes, because this placed an emphasis on the relationship between emotion expression and goal orientation. For example, when predicting student attrition in an online course SA was used in

conjunction with other measures to generate two predictive algorithms which reported a Kappa statistic of 0.403 and 0.432 when predicting attrition (Chaplot et al., 2015). When predicting completion rates for an online course 32 variables were analyzed and the SA measure was ranked 30th in terms of effect size for predicting course completion ($F=4.566$; $p<0.05$; $\eta^2 = 0.014$) (S. Crossley, Place, Mcnamara, Baker, & York, 2016). Another study which compared SA to course completion rates reported results of survival analysis across three courses and found that individual negativity had a significant hazard ratio with course completion in two of the courses (0.84^{**} ; 1.05^{**}) and a significant relationship between individual positivity and completion (1.04^*) (Wen et al., 2014). When comparing SA scores with correct responses in reading comprehension a logistic regression reported a significant relationship ($B=-1.66$, $\chi^2 = 13.8$, $p<0.05$, Odds Ratio=0.87) (Hillaire et al., 2018). Finally, researcher examined if emotion expressed in text in online question and answer boards resulted in the successful outcomes and found that successful posts were identified as negative 23.91% of the time while unsuccessful posts were negative 30.7 percent of the time (Wyner et al., 2008).

Table 2.1.2.1 Emotional theory inferred by measures of accuracy of sentiment analysis in the context of learning

Text	Study	SA Method	BET	SAT	CTE	None
<i>Diaries</i>	(Munezero et al., 2013)	Lexical	-	-	-	+
<i>Evaluations</i>	(Jagtap & Dhotre, 2014)	Machine Learning	-	-	-	+
	(R. A. Calvo & Kim, 2010)	Machine Learning	-	-	+	-
	(Rajput et al., 2016)	Lexical	+	-	+	-
<i>Facebook</i>	(Ortigosa et al., 2014)	Lexical	+	-	-	-
	(Troussas, Virvou, Espinosa, Llaguno, & Caro, 2013)	Machine Learning	+	-	-	-
<i>Forum</i>	(Chaplot et al., 2015)	Lexical	-	+	-	-
	(S. Crossley, Paquette, et al., 2016)	Lexical	-	+	-	-
	(Wen et al., 2014)	Lexical	-	+	-	-
	(Hillaire, Rienties, et al., 2018)	Lexical	+	+	-	-
	(Wyner, Shaw, Kim, Li, & Kim, 2008)	Machine Learning	-	+	-	-
	(Shapiro et al., 2017)	Lexical	-	-	-	+
	(Chang, Maheswaran, Kim, & Zhu, 2013)	Lexical	-	-	-	+
	(Kagklis, Karatrantou, Tantoula, Panagiotakopoulos, & Verykios, 2015)	Lexical	-	-	-	+
<i>Self-Report</i>	(Santos et al., 2013)	Lexical	+	-	+	-
<i>Total</i>			5/15	5/15	3/15	5/15

- Indicates no evidence of theory in measures of accuracy

+ Indicates evidence of theory in measures of accuracy

Finally, when researchers use student perceptions to identify emotion expression this most closely aligns with a CTE perspective because it considers members of the social context as the best candidates to identify accuracy. The CTE suggests that accuracy is not the correct word to use as a social reality can at best be measured in terms of consensus when evaluating SA. Three studies used ratings provided by students in conjunction with SA to compare how SA measures aligned with student ratings. For example, one study compares student of a course evaluations on a scale from 1-5 with sentiment ratings for five algorithms and reported the best classifier in terms of overall accuracy using macro averages as DIM (precision=0.404, recall=0.389, and F-Score 0.363) (Calvo & Kim, 2010). Another study which compared Likert ratings to sentiment scores focused on teacher evaluations and found a correlation of 0.64 between the Likert ratings and the SA score (Rajput et al., 2016). When considering emotions during maths exercises one study asked participants to rate their emotional experience after the activity and write a brief description of their emotional experience. When comparing SA results from analyzing the description with the Likert scores they found a prediction rate of 63% (Santos et al., 2013).

Ignoring accuracy is another approach adopted by researchers that use SA in the context of learning. Five out of fifteen studies examined (33%) used SA but never reported a test for accuracy (Chang, Maheswaran, Kim, & Zhu, 2013; Jagtap & Dhotre, 2014; Kagklis et al., 2015; Munezero et al., 2013; Shapiro et al., 2017). When reviewing 15 studies of SA in the context of AL, I found that only three studies (Calvo & Kim, 2010; Rajput et al., 2016; Santos et al., 2013) evaluated the accuracy of their proposed measure based on the perspective of authors (students). None of the studies used the perspective of the intended audience (see Table 2.1.3.1).

2.1.2.2 IMPLICATIONS FROM EMOTION THEORY FOR EMOTION LEARNING ANALYTICS

In summary, existing research on SA on AL have used a variety of methods to determine accuracy. While often these studies do not explicitly state their theoretical perspectives on emotion the methods used can be mapped to theory, some researchers use multiple methods that are best described by different theories, while other

researchers do not check the accuracy of the measure at all. As a variety of statistical methods were used across the studies the analysis is not conducive to comparing which method produces the best results. However, the results indicate that different researchers based their judgement of accuracy on theoretically different approaches. The emphasis in this PhD thesis is on a CTE approach so it is important to note that none of the methods previously employed by researchers best described by this theory asked students to read their own text and identify the emotion expression that was present. Researchers more frequently relied on a Likert scale rating that was produced in conjunction with the text to determine what the student thought. In the case of a teacher evaluation or course evaluation this might be problematic as the rating provided may or may not have a direct relationship with text provided in those reviews. The gap identified in none of the studies asking students to identify the emotion expression in their own text communication is explored in depth in section 3.3.2.1.1.

2.1.3 HOW SENTIMENT ANALYSIS MEASURES EMOTION USING VALENCE

In this section I detail valence to better understand how researchers using SA in AL define their approaches. Then, I will articulate the bipolar perspective on valence, the bivariate perspective on valence, and the perspective that valence is both bipolar and bivariate. I also consider previous research to understand why different perspectives have developed over the years. Then I consider how previous SA work in the context of learning has measured valence by identifying if previous work detects positive, negative, neutral, and/or mixed valence.

2.1.3.1 WHAT IS VALENCE?

Valence is a dimensional perspective on organizing emotions as positive and negative. There are three competing perspectives on how valence should be organized, as will be described in the three subsequent sections. The first is the bipolar perspective which considers positive and negative to be the opposite ends of the same spectrum. From the bipolar perspective valence is one dimension that can be used to organize emotion (see section 2.1.3.1). For example, the emotion happy can be considered to be

placed on the positive end of the spectrum and the emotion sad can be placed on the negative end of the spectrum. When using a bipolar perspective, emotions are categorized as negative (on the left end of the dimension), neutral (near the center), or positive (on the right). One criticism on the bipolar perspective is that it prevents categorizing emotion as both positive and negative.

When considering the possibility that emotions can be both positive and negative, the bi-variate approach suggests a co-activation where emotions can be categorized as simultaneously activating positive and negative (see section 2.1.3.2). In the bi-variate model there are two variables (one for positive and one for negative). A third point of view, called the evaluative space model (ESM), suggests that emotions are both bipolar and bi-variate. Effectively, the ESM argues that valence should be thought of as a plane. The Y-axis of the plane ranges from neutral to negative and the X-axis of the plane to range from neutral to positive. Points on the X-axis and Y-axis represents bipolar categories of emotion. Points in the X-Y plane represent bi-variate categories of emotion. Many of the arguments between these perspectives are based in critical analysis of emotional measures. To unpack the debates between the models it is best to examine how these three models relate to emotional measures.

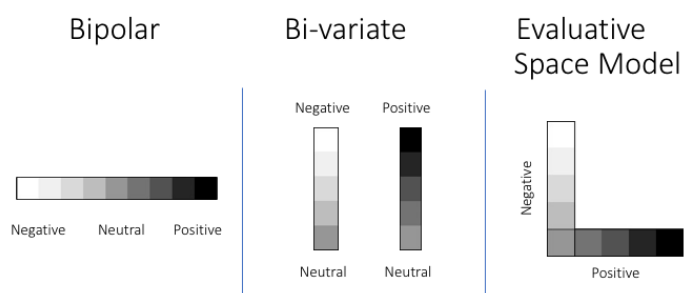


Figure 2.1.3.1 Three Models of Valence: Bipolar, Bivariate, and Evaluative Space

2.1.3.1.1 BIPOLAR PERSPECTIVE ON VALENCE

Bipolar is defined as a reciprocal relationship such that a person experiences positive they necessarily experience less negative, and vice versa (Cacioppo et al., 1999). From the bipolar perspective Russell proposed the model of core affect, which suggests

positive and negative should be used as a bipolar dimension of valence and arousal, with sometimes the addition of the third dimension of control (Feldman Barrett & Russell, 1998; Russell & Barrett, 1999b). Advocates of bipolarity often illustrate the utility of the approach as it appears to produce results (Russell & Carroll, 1999). For example, when testing to see if 296 adjectives could be mapped to a bipolar dimension Mosier (1941) worked with 296 undergraduate students in North America. Mosier's study asked students to rate adjectives on a bipolar scale from 1 to 11 and then checked to see if the distribution of ratings had a normal distribution on the scale. The instructions indicated that "1 means most unfavorable, 6 means neither favorable nor unfavorable, 11 means most favorable, and 0 means cannot be rated" (Mosier, 1941).

There were three exceptions to this hypothesis of normal distributions in the results. The first was that 26 of the terms were identified by more than 10 participants as words they could not rate because they did not know the meaning of words such as Propitious, Cloying, and Iniquitous (Mosier, 1941). The second exception to the hypothesis were skewed distributions which were labelled as "precipice effect". An example of the precipice effect is the term "unnecessary" which had a distribution that was skewed and did not cross the neutral point. This pattern of rating could indicate that some terms are on a univariate scale and not a bipolar scale (Mosier, 1941).

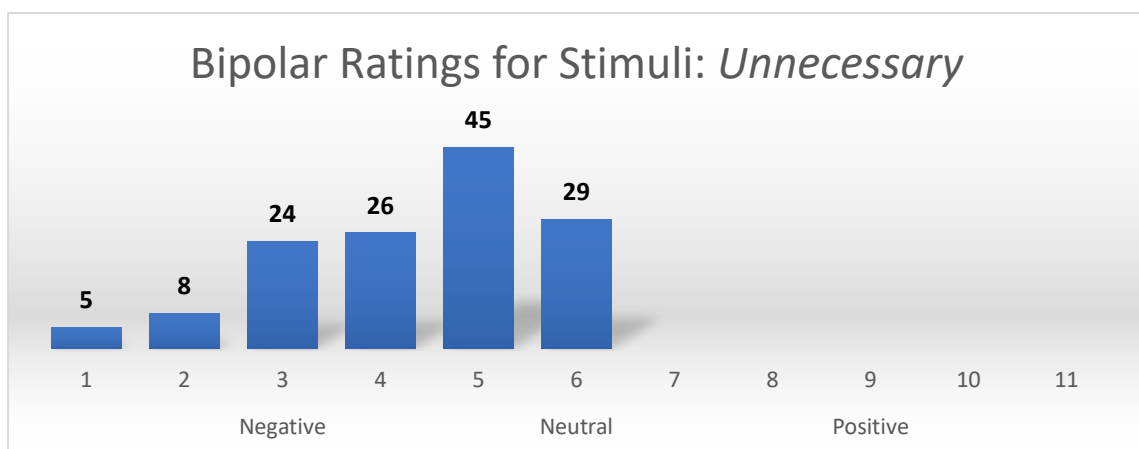


Figure 2.1.3.1.1a Precipice Effect for Stimuli - Unnecessary

The third exception to the hypothesis was distributions with two or three peaks. The two peak distributions had one peak at the neutral point and the second peak was either positive or negative. The three peak distributions had a peak at positive, neutral, and negative. These distributions were labelled "bimodality of meaning". An example of the

“precipice effect” is the stimulus “Completely Indifferent”. This bimodality of meaning indicates that the stimulus can have multiple interpretations. While some evaluate the stimuli, they make an interpretation of one possible category of positive, negative, or neutral. Given the multiple modes this could indicate that the stimulus has aspects that in multiple categories. The bimodality of meaning was present in 28 terms or roughly 9% of the stimuli.

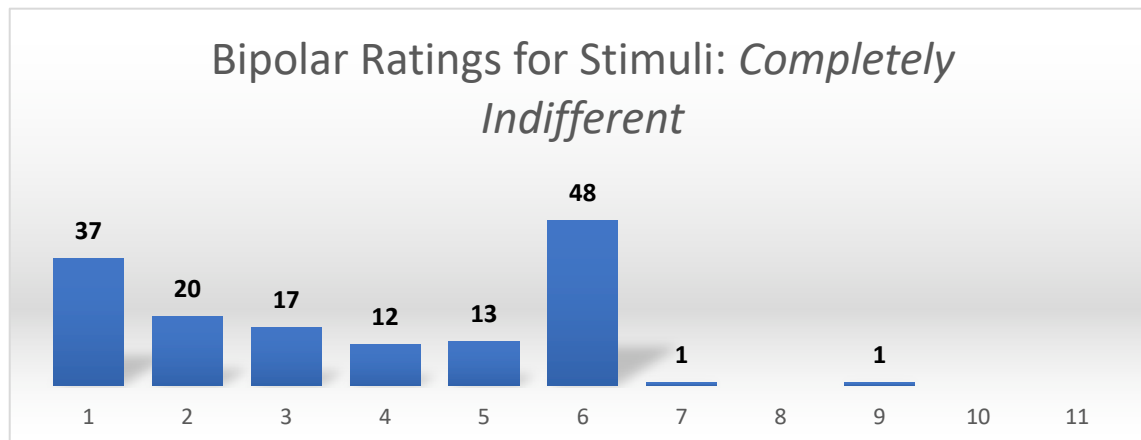


Figure 2.1.3.1.1b - Bimodality Effect for Stimuli - Completely Indifferent

The majority 218/296 (73.65%) of stimuli had linearity indicated by the normal distribution. The remaining cases 78/296 (26.35%) could be explained through further analysis. For 24/296 (8.10%) of the stimuli were so close to linear that the only non-linearity was found at the extremes of the scale and could be considered to map to the bipolar scale. For 25/296 (8.45%) the non-linearity can be attributed to participants being unfamiliar with the term indicated by more than 10 participants rating the stimuli as “0 cannot be rated”. For 19/296 (6.42%) of the non-linearity indicated bimodality of the distribution. 4/296 (1.35%) had a precipice effect. This left 6/296 (2.03%) stimuli (i.e. Exasperating, Invigorating, Ordinary, Piteous, Pitiable, and Tragic) without an explanation. From these results I could consider this as evidence that valence as a dimension is perhaps mostly bipolar with 242/296 (81.76%) exhibiting linearity or close to linearity. With insufficient evidence to explain 31/296 (10.47%) of the stimuli as either unfamiliar terms or lacking a structural pattern that has a logical explanation. For the remaining 23/296 (7.77%) stimuli evidence that positive and negative may be best described as their own dimension.

It is possible that asking people to rate words on a bipolar scale resulted in this bimodality of meaning because there was a co-activation of positive and negative response. Raters that focused on the negative aspects may have provided negative ratings while those focusing on the positive aspects provided a positive rating. Some may have considered both positive and negative aspects to cancel each other out resulting in a neutral rating. To better understand the co-occurrence of positive and negative valence I next examine the bivariate perspective on valence. To take insight from the analysis conducted by Mosier (1941) I consider alternatives to the bipolar approach towards valence and consider how text may have a bimodality of meaning (i.e., the same text may have both positive and negative interpretations). In fact, it is possible that someone who reads text that has a bimodality of meaning that they might report that it is both positive or negative. However, given a bipolar scale to rate such a term has demonstrated that it results in a bimodal distribution.

2.1.3.1.2 BIVARIATE PERSPECTIVE ON VALENCE

One of the broadly used bivariate valence psychological measures is the Positive and Negative Affect Schedule (PANAS) (Leue & Beauducel, 2011). PANAS asks participants to rate on a Likert-scale how much their experience is described by 10 positive words and 10 negative words (Watson, Clark, & Tellegen, 1988). Then the positive affect scores are averaged and the negative affect scores are averaged. This results in two values a positive score and a negative score which are considered orthogonal dimensions (Watson et al., 1988).

There are debates that take strong positions that valence is best measured as bipolar or bivariate (Green et al., 1993; Russell & Carroll, 1999; Watson et al., 1988). There is an argument that valence represents an evaluative space that includes both a bipolar dimension as well as the two-dimensional plane of positivity and negativity (Cacioppo et al., 1999; Cacioppo, Larsen, Smith, & Bertson, 2004). There are a number theoretical approaches to considering how the values on the plane of positivity and negativity would map to a dimension of mixed valence (Kreibig & Gross, 2017). The dimension of mixed valence will be critically examined in section 2.1.3.3. This section with focus on the valence debate of bipolarity, bivariate, or both. Before taking a position on this I can

examine empirical studies for evidence of bipolarity and bivariate in language expression. First by examining short word stimuli such as single words or phrases. Second examining literature on text message processing for empirical evidence.

While I can see evidence of short stimuli that are both bipolar and bivariate, the issue becomes a little more complex with text that is longer such as text messages. For example, what should be done if a sentence has a word that is positive and word that is negative (“I had a positive reaction to what u wrote [Student_10] because I have not a clue what the Null hypothesis is”). In this example from a chat interaction between students from one of studies in this PhD, a student indicated that he had a positive reaction to a message from Student_10, but at the same time he did not really know how to answer the main question of the hypothesis in question. Whether or not this is a positive, negative, or mixed emotion of course can be debated.

Therefore, given the focus of this thesis on detecting emotion in written text using SA it is important to briefly define the lexical approach towards SA, where a dictionary is commonly used to score the positive and negative aspects of a text message. When using a lexical approach for SA there is a dictionary which has both a list of words and associated scores. For example, the Warriner et al. dictionary (Warriner, Kuperman, & Brysbaert, 2013) uses valence scores that range between 1 and 9 for the words in the dictionary. Values below 4 are considered negative words, values above 6 are considered positive, and values between 4 and 6 are considered neutral (Warriner et al., 2013). When a phrase contains both a negative and a positive word it is necessary to make an interpretation. For example, the SA technology SentiStrength produces two ratings – one positive and one negative. Each rating is on a scale from 1-5. If the SentiStrength produced a rating of positive: 5 and negative: 5 then the message would have been predicted to contain both positive and negative. However, when interpreting this prediction there are a variety of methods used in SA research.

One approach is to consider the position of positive and negative elements of the communication. In contexts such as movie reviews the whole does not appear to be the sum of the parts as many consider the phenomena of “thwarted expectations”, where a review contains both positive and negative statements, and the final viewpoint expressed is an overall viewpoint (Pang, Lee, & Vaithyanathan, 2002; Turney, 2002). Some

researchers in education have adopted the strategy to preference the end of communication as more important (Ortigosa et al., 2014). In other contexts like automobile reviews the whole of the review does appear to be the sum of the parts (Turney, 2002). This indicates that in some contexts I can simply identify the positive and negative elements and make an interpretation of the whole communication by calculating an overall score based on scores associated with parts of the message (Turney, 2002). When taking this approach this raises questions about the best approach to computing an overall score based on the positive and negative aspects of the communication.

The simplest approach toward computing an overall score indicated by Hershfield & Larsen's (2012) review of mixed emotion measures is a dichotomous co-occurrence index, which is a binary indication that there is both positive and negative present. The dichotomous co-occurrence index has been criticized for over predicting mixed emotions (Hershfield & Larsen, 2012). A method, which does not have a specific label and I will call "mixification", is nearly identical to neutralization where positive and negative scores are added up and when positive and negative scores are equal it indicates a mixed statement (Rajput et al., 2016). An alternative strategy to quantify a mixed score is to calculate the absolute difference between the positive and negative score (Hui, Fok, & Bond, 2009). A favored approach in SA seems to be to quantify mixed emotion by taking the MIN score (Hershfield & Larsen, 2012), which is defined as taking the minimum score between the positive and negative scores. The positive and negative scores must first be placed on comparable scales (e.g., positive rated from 0-1 and negative rated from 0-1) before calculating the MIN score (Hershfield & Larsen, 2012). Table 2.1.3.1.2 illustrates five approaches toward interpreting positive and negative measures using neutralization, mixification, ABS, and MIN strategies to illustrate the characteristics of the approaches. I have provided five bivariate scores to illustrate how the five methods differ in their interpretation which are in Table 2.1.3.1.2.

Table 2.1.3.1.2 Five bivariate scores interpreted by five methods to compute mixed emotion expression

Score	Neutralization	Mixification	ABS	MIN	Co-Occurrence
Positive – 4 Negative – 4	Neutral (0)	Mixed (0)	Neutral (0)	Mixed (4)	Mixed
Positive – 4 Negative – 3	Positive (1)	Positive (1)	Mixed (1)	Mixed (3)	Mixed
Positive – 4 Negative – 1	Positive (3)	Positive (3)	Mixed (3)	Mixed (2)	Mixed
Positive – 1 Negative – 1	Neutral (0)	Mixed (1)	Neutral (0)	Mixed (1)	Mixed
Positive – 1 Negative - 0	Positive (1)	Positive (1)	Positive (1)	Positive (1)	Positive

If you review Table 2.1.3.1.2 ABS indicates a greater degree of mixed when the positive and negative ratings are further apart. The MIN score indicates a greater degree of mixed as both positive and negative ratings are higher. By taking all possible combinations of a bivariate measure, which places positive and negative scores with a rating from one to five for each dimension of positive and negative, with one representing neutral, and values two through five representing increased presence of the dimension, I can see how five approaches towards interpretation interpret all possible values. I can arrange the methods as favoring neutral interpretations to favoring mixed interpretations as follows: Neutralization, Mixification, MIN, ABS, Co-occurrence. Figure 2.1.3.1.2 shows a bar chart of how all possible combinations result in this classification.

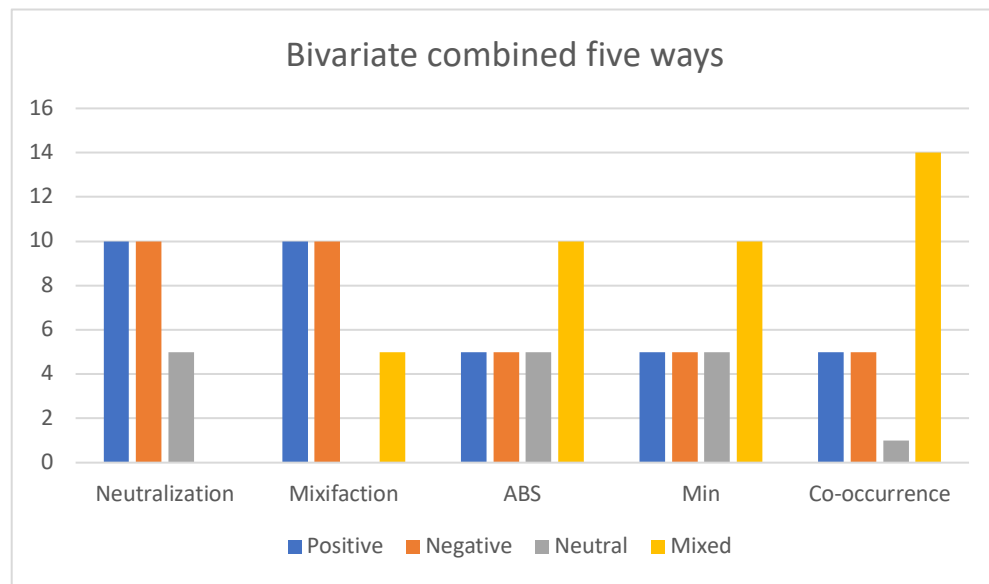


Figure 2.1.3.1.2 All possible bivariate rating with positive and negative scales from 1-5 categorized

A review of these strategies illustrates that the MIN approach appears to be the favored method to interpret mixed emotion (Hershfield & Larsen, 2012). However, it is hard to say which strategy would be the best. Based on the simulated results the co-occurrence strategy classifies the most text that has both positive and negative as mixed.

2.1.3.1.3 VALENCE AS BOTH BIPOLAR AND BIVARIATE

One way to make sense of the distinction between dichotomous co-occurrence and the three scaling strategies (i.e. Neutralization, Mixification, ABS, MIN) is to ask if mixed emotion is like neutrality, which is viewed as a binary classification, or if mixed is like positive and negative which are viewed as scaled dimension(s). By stepping out of the context of SA which has a focus in the literature of fitting to the review context on the topic (Pang et al., 2002; Turney, 2002), and examining the broader context of all measures of mixed emotion (Hershfield & Larsen, 2012; Hui et al., 2009), this section outlined challenges that arise from interpreting a bivariate measure (positive and negative as two independently reported dimensions) to infer mixed emotion. To take

another step back and make the problem more abstract I can examine paradigms of mixed emotional measurement.

2.1.3.1.4 BIVARIATE, UNIVARIATE, AND MULTIVARIATE MIXED PARADIGMS

When considering how to model valence that includes mixed emotion it is necessary to determine which paradigm of measurement is the most appropriate. There are three paradigms to consider. The first paradigm is subjective bivariate mixed emotion, which is a measure where positive and negative are reported separately, and mapped into a univariate dimension of mixed emotion (Kreibig & Gross, 2017). When taking this approach it leads to interpreting a positive and a negative score in order to infer mixed, which can lead to five strategies of interpretation: neutralization, mixification, abs, min, and co-occurrence (Hershfield & Larsen, 2012). The second paradigm of subjective univariate mixed emotion (UME) is a measure which directly asks the subject to report the category of positive, negative, neutral, and mixed emotions (Kreibig & Gross, 2017). While this strategy does not require the inference of mixed from a positive and negative measure it does require that the interpretation of what qualifies as mixed. The third paradigm is subjective multivariate mixed emotion where multiple measures of positive and negative are interpreted as mixed when there are positive and negative values across the multiple measures. If across multiple measures there are purely positive or purely negative responses the subjective multivariate mixed measure can be mapped to the bipolar dimension of valence (Kreibig & Gross, 2017).

Given the position articulated in section 2.1.3.2 identified that the existence of both positive and negative can be interpreted in a bivariate manner, which would be considered mixed and could also be interpreted in a bipolar manner, then using a subjective bivariate mixed measure would require an interpretation as to whether the positive and negative valence reported should be interpreted through neutralizations or mapped into the univariate mixed dimension using one of the possible strategies. Alternatively, by taking a subjective univariate mixed approach in cases where the both positive and negative are perceived, the subject would make the interpretation as to whether the positive and negative dimensions should be mapped to the univariate mixed

dimension or integrated (e.g. neutralized - see section 2.1.3.2) into the bipolar dimension. When creating a measure, I adopt the univariate mixed paradigm. By asking students to label messages as positive, negative, neutral or mixed (see section 3.2.2.1) I avoid inferring mixed based on two measures (positive and negative) and rely on the interpretation of students to select the best category.

2.1.3.2 SENTIMENT ANALYSIS METHODS

To review existing work on sentiment analysis I organize previous work along three dimensions. First, I examine how sentiment analysis establishes *ground truth* which is defined as what sentiment analysis researchers consider the correct sentiment label for text. Second, I consider the unit of analysis that previous work used when analyzing text. Third, I consider techniques used by sentiment analysis researchers.

2.1.3.2.1 GROUND TRUTH AND SENTIMENT ANALYSIS

To evaluate the accuracy of sentiment analysis the first step is establishing ground truth which is what is considered to be the correct label for text. Recent work critically examining the process of establishing ground truth highlighted that the approach used to establish ground truth is itself a design task which is under-evaluated (Muller et al., 2021). Muller (2021) states that while many researchers take the first step of establishing ground truth that once they complete this task the process “fades into the background” and the method used to establish ground truth becomes adopted as objectively correct even though frequently ground truth was established by humans making decisions about what is correct. While Muller (2021) was criticizing the practice in establishing ground truth for computer vision classifiers the conceptual concerns are the same when establishing ground truth for sentiment analysis.

In this thesis I take the perspective of CTE which defines emotion as a social consensus. It is from this theoretical foundation that I design my method to establish ground truth by building on existing work in crowdsourcing ground truth. Crowdsourcing is a method that establishes ground truth by seeking to benefit from the so called “wisdom of the crowd” where many people judge stimuli such as text to determine an appropriate label (Morris & McDuff, 2009). While using multiple ratings

from a crowd fulfills the social consensus aspect of CTE it still lacks the contextual awareness (recall the example Barrett provides of misinterpreting the expression of Venus Williams as anger when it was more likely elation in section 3.1). To align crowdsourcing with CTE it would necessitate that the crowd is from the social context where the text is generated. Crowdsourcing described both crowds labeling existing data and generating new data (Morris & McDuff, 2009). To build on this term, I refer to students generating text during online chat conversations and providing labels of their own text as *student sourcing*. Student sourcing is effectively crowdsourcing in the context of learning where students both generate the text in discussions where the same students in those discussions also provide valence labels. In taking this approach I consider students to be subject matter experts of their own discussions because the opinions of the students are either the opinions of the author or the intended audience which by definition is the goal of SA. To devise an approach to student source labels we next review literature on crowdsourcing.

The adoption of services such as Mechanical Turk to conduct research is widespread. Evaluations of crowdsourcing platforms such as Mechanical Turk for single label tasks such as sentiment analysis (Zheng, Li, Li, Shan, & Cheng, 2016) suggests that the gold standard is 20 raters on the platform with majority vote determining ground truth, but in cases where it is not feasible to get 20 raters on every text the suggested alternative approach for high quality outcomes is to use fewer raters and apply the expectation maximization algorithm proposed by Dawid and Skeene (1979) which computes majority vote and iterates through to evaluate the quality of each rater and updates selection of ground truth considering rater quality. One of the explanations for why it is not always feasible to get 20 raters for every message is that the task of labeling can be tedious and time consuming (Morris & McDuff, 2009; Raykar et al., 2010). When analyzing the quantity of ratings there is evidence that single ratings (e.g., text with a single rating) from non-experts can be detrimental to the outcome (Hsueh, Melville, & Sindhvani, 2009). At the same time there is evidence that it is better to label many records once than fewer records multiple times when worker quality is good (Khetan, Lipton, & Anandkumar, 2017).

Based on this review of crowdsourcing as an approach the first question to ask is: what quality rating can we expect from students? A study at a Scottish University compared college student ratings to expert raters for specific emotions and found that students had a reasonable level of agreement with expert raters (Gill, Gergle, French, & Oberlander, 2008) which was better when considering text over 200 words and less accurate when considering text less than 50 words from blog posts. However, they noted that the length of text did not factor in when the emotions strongly expressed valence. To frame these results in terms of the CTE, when students rate the valence of text there is evidence that there is a capacity to do so as compared with what researchers perceive as emotion in text. Section 2.1.2.1 identified a gap in having students identify emotion in text and Gill et al. (2008) found that students have the capacity to identify valence in text. While agreement with researchers is how Gill et al. (2008) evaluated student accuracy from the CTE perspective on emotion unless the researchers were part of the University community disagreement between students and researchers would not be considered inaccuracy of student ratings. The strongest argument to consider students expert raters from CTE is that to be an expert rater it is necessary to be a part of the social context. While there is a theoretical, and some empirical evidence that students might be high quality raters there is a need to consider how to evaluate their accuracy. In the next section I propose two forms of ground truth used in this thesis.

While I consider students experts it is also important to report the level of agreement when reporting the results of crowdsourcing annotations. In social computing two common metrics to report are Krippendorff's Alpha and Fleiss' Kappa (Salminen, Al-Merekhi, Dey, & Jansen, 2018). Krippendorff's alpha is an agreement statistic which is applicable to 1) any number of values per variable, 2) any number of raters, 3) small and large sample sizes, 3) multiple metrics including nominal data, and 4) data with missing values (Krippendorff, 1980, pg. 221). Krippendorff's alpha is formally defined as $1 - (\text{observed disagreement} / \text{expected disagreement})$ where expected disagreement is based on chance levels of disagreement. While it is important to report agreement statistics there are suggestions that the results can be misleading in social computing because low agreement statistics might simply indicate that there are differences of opinion (Salminen et al., 2018). In a systematic review of crowdsourcing in social computing

agreement statistics averaged 0.60 which is lower than typical threshold values (Salminen et al., 2018). When using crowdsourcing for highly subjective topics such as what text raters find “interesting” an evaluation of crowdsourcing averaged a Krippendorff’s alpha score of 0.01 (Alonso, Marshall, & Najork, 2013). Fortunately, SA is not considered to be this hard to agree upon. To establish expectations, I next review studies that consider SA using the four valence categories of positive, negative, neutral, and mixed. Krippendorff provides a conservative interpretation suggesting that alpha values less than 0.67 should be disregarded, tentatively evaluations for values between 0.67 and 0.80, and conclusions can be made for values above 0.80 (Hallgren, 2012). While this conservative approach has merit the field of social science mean score of 0.60 suggests a less conservative approach is warranted so we adopt the agreement statistics interpretation from Landis and Koch (1977) which suggests 0.0 to 0.2 slight agreement, 0.21 to 0.40 fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement, and 0.81 to 1.0 almost perfect or perfect agreement.

In previous sentiment analysis work that examined sentiment analysis considering the binary classification of positive and negative as compared with the differentiated classification which included neutral and mixed valence the binary classification with five raters had a Krippendorff’s alpha of 0.47 (moderate agreement) while the differentiated agreement was 0.22 (fair agreement) and in this study the binary classification was adopted based on moderate levels of agreement (Schmidt & Burghardt, 2018). With two annotators labeling the categories of positive, negative, neutral, and both much higher agreement was achieved reporting agreement at 82%, but agreement for the category of both was only 50% and only 6 out of 447 items analyzed were labeled as both by either of the raters (Wilson, Wiebe, & Hoffman, 2005). Using eleven raters categorizing messages as positive, negative, neutral, and mixed with nominal categorization the agreement reported was Krippendorff’s alpha of 0.65 (substantial agreement) and they further analyzed the ratings in an ordinal approach which considered disagreement of positive and negative to be more severe they computed Krippendorff’s alpha of 0.68 (substantial agreement) (Chakravarthi, Muralidaran, Priyadharshini, & McCrae, 2020). As we see alpha scores reported of 0.22 and 0.65 when rating valence into four categories there is clearly some subjective

judgement in play, but it is nowhere near as bad as the highly subjective scores of 0.01 (Alonso et al., 2013). Based on this review I consider moderate agreement in the range of 0.41 to 0.60 as the cutoff when evaluating the work in my thesis to indicate there is sufficient agreement.

2.1.3.2.2 GROUND TRUTH USED IN THIS THESIS

To respond to the criticism that frequently crowd sourcing methods considers the crowd anonymous and disconnected from the context I examine a crowd of students to label their own messages. To respond to the criticism that training raters forces homogeneity of perspectives I minimize instructions for the raters. The approach of minimal instructions and students rating their own messages seeks to benefit from the intuition of the students evaluating their own communication (Waldinger, Hauser, Schulz, Allen, & Crowell, 2004). One challenge with having students evaluating their own communication is that the task can be tedious. Rather than having students evaluate every single message from their discussions I instead ask them to provide 1-3 examples for each valence class. This inevitably generates many messages with single ratings. As I adopt the perspective that students are experts in terms of evaluating their own group discussions single ratings for more messages are preferable over multiple ratings on fewer messages.

As this thesis introduces the student sourcing approach, crowd sourcing with students, I also follow the standard practice of using Mechanical Turk to crowd source ratings as a comparison the student sourcing providing an external perspective to evaluate if student opinions generate a ground truth that is both different and useful as compared with external perspectives.

2.1.3.2.3 SENTIMENT ANALYSIS METHODS

As mentioned earlier a common definition of Sentiment Analysis is that it aims to detect the opinion of the author of the text and the reaction elicited by the intended audience of the text. More specifically, SA can both predict specific emotions or valence of text. In this thesis I focus on sentiment analysis designed to detect the valence of text.

Lexical based approaches for SA create a dictionary of terms frequently associated with each valence class (e.g., positive, negative) and the presence of words from the dictionary in text are used to predict which valence class to classify the text.

Machine Learning approaches frequently used supervised learning where a set of messages have been labeled with ground truth (see section 2.1.3.2) valence classes and then features of the text are extracted to train a classifier based on the ground truth labels. There are many machine learning classifiers used to train SA classifiers. A study suggested that Naïve Bayes (NB), and Support Vector Machines (SVM) outperformed most algorithms for SA (Altrabsheh et al., 2013). SVM effectively attempt to identify a hyperplane which separates different categories of data by using a margin by drawing a vector across the data points at the margin to classify text based on features (Shawe-Taylor & Cristianini, 2000). NB uses data available as evidence that text belongs to a specific class by predicting the probability that the text belongs in the class based on the available features (Leung, 2007). Logistic regression (LR) is a third method frequently used to train a sentiment analysis classifier which uses a sigmoid function and establishes coefficients for all features used to train the classifier generating a binary classification. SVM, NB, and LR are three common approaches in machine learning for SA.

While there are also more recent developments in terms of deep learning such as LSTM, RNN, and CNN which are frequently used to train classifiers while deep learning algorithms can generate higher levels of accuracy the models they generate are less transparent as it is difficult to interpret them. In this thesis I use logistic regression to train a sentiment analysis classifier as the model is easier to interpret.

2.1.3.2.4 SENTIMENT ANALYSIS AND DOMAIN DEPENDENCY

From the early days of sentiment analysis research, the domain has been considered a critical element. A classifier trained on movie reviews is more likely to be accurate when used on similar data such as movie reviews or possibly on other review data such as product reviews. Effectively the context where the classifier is trained in some manners dictates which contexts it will be the most effective.

Research on the use of sentiment analysis in education have pointed out that the context is problem is amplified as the domain for any given course is both considered in the domain of the subject being taught and the fact that the context is one where students are learning. For example, the context of a statistics course would have a very different interpretation of the word ‘mean’ as there is a connotation to the word mean as a type of average. In a course on say art the word mean might more likely indicate unkindness. The fact that learning is itself a context is also a factor in that students may not fully understand the meaning of the words they are using (e.g., earlier drafts of this thesis called into question my understanding of some of the terminology) and that lack of understanding can lead to using terms inappropriately (which I hope I have addressed in this version of my thesis).

When considering how to make sentiment analysis for a specific domain it requires getting a small sample of data from the domain, labeling it, and using it to train classifiers (Yadollahi, Shahraki, & Zaiane, 2017). Due to the context sensitive nature of SA as a technology the pursuit in this thesis to train a classifier using student data is justified as none of the general technologies used were trained with student data. It is for this reason I anticipate that the classifier trained in the context of learning should outperform general measures built to solve general purpose problems in SA. While I expect the classifier trained with student data will outperform general SA technologies used on student data the goal of this thesis is not to build a measure for general purposes as that would actually be in stark contrast with the theoretical perspective of CTE. so, it would likely perform much worse than more established general measures in other contexts.

2.1.3.2.5 MAPPING VALENCE THEORY TO SENTIMENT ANALYSIS STUDIES IN THE DOMAIN OF LEARNING

When considering valence categories measured when applying SA to the context of learning, there appears to be an emphasis in the existing literature on measuring positive and negative valence. Of the 15 SA of AL studies reviewed, all of the studies measured both positive and negative valence (see Table 2.1.3.2.5). About half of the studies, 7 out of 15, measured the category of neutral, and only 2 out of 15 studies measured a

category of mixed emotion. Only one study (Santos et al., 2013) measured all four categories of positive, negative, neutral, and mixed. However, they referred to mixed as ambivalence which they defined as both positive and negative, and not enough attention was paid to measuring neutral and mixed expression. In the second study (Rajput et al., 2016) that measured positive, negative and mixed expression the authors used neutral and mixed interchangeably when describing the results but reported statistics for the category of mixed expression.

This thesis adopts the univariate mixed paradigm (which measures positive, negative, neutral, and mixed) (Kreibig & Gross, 2017). By measuring positive, negative, neutral, and mixed I explore the gap in the literature identified by Santos et al. (2013) that not enough attention is paid to neutral and mixed expression. To detect mixed expression both Rajput et al. (2016) and Santos et al. (2013) used a co-occurrence method to interpret bivariate measures which means they detect positive and negative separately and categorize expressions that are both positive and negative as mixed.

Table 2.1.3.2.5 Valence categories for SA used in the context of learning

Text Source	Study	Positive	Negative	Neutral	Mixed
Evaluations	(Munezero et al., 2013)	+	+	-	-
	(Jagtap & Dhotre, 2014)	+	+	-	-
	(R. A. Calvo & Kim, 2010)	+	+	+	-
Facebook	(Rajput et al., 2016)	+	+	-	+
	(Ortigosa et al., 2014)	+	+	+	-
Forum	(Troussas, Virvou, Espinosa, Llaguno, & Caro, 2013)	+	+	-	-
	(Chaplot et al., 2015)	+	+	+	-
	(S. Crossley, Paquette, et al., 2016)	+	+	-	-
	(Wen et al., 2014)	+	+	-	-
	(Hillaire, Rienties, et al., 2018)	+	+	+	-
	(Wyner, Shaw, Kim, Li, & Kim, 2008)	+	+	-	-
	(Shapiro et al., 2017)	+	+	+	-
	(Chang, Maheswaran, Kim, & Zhu, 2013)	+	+	-	-
	(Santos et al., 2013)	+	+	+	+
Self-Report	(Kagklis, Karatrantou, Tantoula, Panagiotakopoulos, & Verykios, 2015)	+	+	+	-
	(Santos et al., 2013)	+	+	+	+
Total		15/15	15/15	7/15	2/15
- Indicates no detection of valence category					
+ Indicates detection of valence category					

2.1.4 EMOTIONAL MEASURES BEYOND VERBALIZATION AND COMMUNICATION

While the emphasis of this thesis is the creation and validation of a SA technology in the context of learning, an important part of the validation process is to select emotional measures that can be used to validate the proposed SA measure. I organize self-report of emotion into two categories: state and trait, whereby I will distinguish four distinct emotion measurements, namely React, PANAS, BEQ, and MES in the next two sections.

2.1.4.1 STATE

Emotional state is defined as momentary experiences of students for which students' self-report are the primary means to access the state of students (Ainley, 2007). The development of emotional tool React (Hillaire, Rappolt-Schlichtmann, & Ducharme, 2016) is in part what inspired this thesis topic. When developing this measure the focus was on a self-report mechanism designed to support students to reflect on their emotional reactions to learning materials (Hillaire et al., 2016) effectively measuring their emotional state. Initial analysis of how React responses related to SA comments in the Udio platform for middle school students in North America suggested there is potential benefit in measuring both self-report of emotion and emotional expression through trace data to better understand the emotional experience of students (Hillaire, 2015). In tandem with the work on this PhD thesis I also found a relationship between React responses and learning outcomes of these middle school students (Hillaire et al., 2018). As the development of the React measure inspired this work it was included as the self-report mechanism for emotion state at the beginning of the lab activity (see section 3.3.3.2). As this measure has shown promise in conjunction with SA measures to investigate learning (Hillaire et al., 2018) it is one of the emotional state measures I use for emotional state.

One of the more ubiquitous measures of emotional state is the PANAS. PANAS is an ideal measure to evaluate the overall experience from a bi-variate perspective because it was designed based on that theoretical perspective (Watson et al., 1988, 1999). However, PANAS was developed considering the parallel perspective on valence (Watson et al., 1988). The ESM proposed by Cacioppo et al. (1999) points out that bi-variate measures are sufficient to measure emotions in some circumstances depending on where participants are in the evaluative process. However, bipolar (integrating positive and negative) approaches can at times do a better job of measuring emotion (Cacioppo et al., 1999; Feldman Barrett & Russell, 1998). As mentioned above when reviewing how emotion expression and experience relate to in the consensual model of emotion all referenced studies used a PANAS scale to measure the experience of emotion (Gross, John, & Richards, 2000; Kahn et al., 2016).

Researchers that have engaged in the debate between bipolar and bivariate measurement approaches towards valence have examined statistical indications of measurement error for scales that calculate affect in both bipolar and bivariate ways (Cacioppo et al., 1999; Green et al., 1993). Specifically, PANAS has been criticized by some researchers in that the scale works by having participants rate 12 distinct emotions words (6 positive and 6 negative) to describe an overall experience. However, the words on the PANAS scale are not polar opposites (e.g., the scale does not include opposite terms like happy and sad), but rather are comprised of terms that are designed to yield independent scores (Cacioppo et al., 1999; Feldman Barrett & Russell, 1998; Green et al., 1993). While this independence of measurement is helpful in measuring the parallel aspect of the ESM (Cacioppo et al., 1999), PANAS is designed specifically to avoid measuring integrative aspects of valence. Part of the challenge with measuring the overall emotional experience is that measures need to model the entire universe of the construct (Weidman et al., 2016). So while PANAS is designed to capture the parallel aspect of valence it does not include the integrative aspect of valence, and both are necessary components of the ESM (Cacioppo et al., 1999).

This perspective is consistent with the general criticism that emotion researchers frequently focus on one aspect of emotions, so disagreements between emotion researchers can plausibly be explained by taking the perspective that those

disagreements stem from focusing on different facets of emotions (Pekrun, 2005). In this thesis the perspective adopted is ESM (Cacioppo et al., 1999) which includes a bi-variate component to valence. Critics have pointed out how the design of the tool generates bi-variate measurement of valence at the expense of not capturing bi-polar or integrative valence data in the self-report (Green et al., 1993).

To fill the gap of measuring the integrative aspects of valence I include the Mixed Emotion Scale (MES), which takes a distinctly different approach towards measuring valence from both a theoretical perspective and from a practical perspective. From a theoretical perspective the MES aims at explicitly measuring the integrative nature of positive and negative valence, and was shown to be distinguished as a measure from the related measures (Berrios & Totterdell, 2013) of ambivalence (Pekrun et al., 2011) and intolerance of ambiguity. The MES does this from a practical perspective by asking participants to rate their agreement with statements that describe both parallel and integrative experiences. Rather than asking participants to describe positive and negative aspects of their experience using ratings for emotion words (like the PANAS does), the MES asks participants to rate their agreement with statements that describe parallel and integrative emotional experiences. For example, the MES asks participants to rate their agreement with the statement “I felt a mixture of emotions” (Berrios & Totterdell, 2013). By directly measuring the experience of mixed emotions the MES is best described as a measure based on the univariate mixed paradigm (Kreibig & Gross, 2017), which is the same paradigm used in the development of SSSAC Logistic (see Study 1 section 4.1).

While there is a theoretical argument to include the MES in this validation the tool was recently developed (Berrios & Totterdell, 2013). Previous work that used the MES focused on how mixed emotion relate to cognitive process such as goal conflict (Berrios, Totterdell, & Kellett, 2015). While there has been some recent exploration to defined mixed emotion (Kreibig & Gross, 2017; Larsen, Coles, & Jordan, 2017), the mixed emotion scale (MES) has not seen widescale adoption in emotion research. Therefore, the purpose for including the MES is based on theoretical alignment rather than previous empirical work.

2.1.4.2 TRAIT

Trait is defined as how a person behaves over a long period of time (Fleeson, 2004). As a model of considering how emotional trait relates to emotional expression one experiment (Gross et al., 2000) was conducted with 76 undergraduate students in North America with an average age of 21 (SD = 3 years), referred to as the target participants. An additional 228 participants were peers who knew at least one of the target participants for three years. For each target participant there were three peers who knew the target participant and rated how frequently the target participants expressed four positive emotions (amusement, joy, love, and pride) as well as ratings for frequency of expressing negative emotions (anger, fear, sadness, and shame). In the study each target participant was administered the Berkley Expressivity Questionnaire (BEQ) as a dispositional measure for emotion expression, which has the subcomponents of positive expressivity and negative expressivity. Furthermore, this study used the PANAS instrument described in the previous section to measure the tendency for the participants to have positive and negative experience. The results indicated that the peer ratings for positive expression had a positive correlation in a moderated multiple regression with both the BEQ for positive expressivity (.32) The peer ratings for negative expression had a positive correlation in a multiple regression with both the BEQ for negative expressivity (.37).

However, another study (Kahn et al., 2016) conducted an experiment with 66 college students where participants were asked to watch a brief film and provide an oral report about their reaction to the film. The words from the oral report were transcribed and the text was analyzed using LIWC (previously described in section 3.3.3.1.6 as a benchmark for SSSAC Logistic). The results indicated that positive emotion detected by LIWC correlated with positive expressivity measured by BEQ with a correlation coefficient of 0.25. Negative expressivity did not correlate with negative emotion detected by LIWC in the same experiment (Kahn et al., 2016). While the Kahn et al. (2016) study had mixed results with transcribed oral responses with the BEQ they conducted another experiment with written expression compared to the LIWC, whereby 79 undergraduate students in North America were instructed to write about a topic that

varied by three conditions. The first condition asked participants to write about a time in their life where they were amused. The second condition asked participants to write about a time in their life when they were sad. The third condition asked participants to describe a typical day. The conducted a series of 3x2 (condition x order) mixed ANOVAs and found the amused condition participants used more positive words ($F(2.76) = 94.69, \eta^2 = .71, p < .001$), while participants in the sad condition used more negative words ($F(2.76) = 105.75, \eta^2 = .74, p < .001$).

2.2 THEORY AND DESIGN OF LEARNING

As the review of emotional theory established that the Constructive Theory of Emotion was ideal for the focus of this PhD thesis I now turn to learning theory. Our emphasis on emotional verbalization and communication illustrated that I can anticipate both self-regulation of emotion through communication and social regulation of emotion through communication which informs our examination of the self and socially regulated learning perspective and how it intersections with emotion. With a foundation of learning theory self and socially regulated learning I then consider how the design of learning environments might support emotional communication by reviewing literature in computer supported collaborative learning (CSCL) and universal design for learning (UDL). As the aim of this PhD thesis focuses on the creation and validation of a SA measure using the Constructive Theory of Emotion as the foundation the purpose for this section is simply to highlight how this measure may be applicable in terms of learning theory and what kinds of learning design might support the development and evaluation of the measure.

2.2.1 LEARNING THEORY AND EMOTION

Self and Socially Regulated Learning (SSRL) considers the implications of the intersection of the self and social context from a regulation perspective in the context of learning. Before thinking about emotion in the context of SSRL It is helpful to first understand the seminal work on self-regulated learning (SRL) and then examine extensions of SRL to accommodate emotions as well as extensions to consider social

context. From that foundation I can connect SRL to both social and emotional extensions. There are three perspectives that are commonly referenced in SRL. Zimmerman, Winne & Hadwin, and Pintrich & De Groot. There are four phases in the Winne & Hadwin (1998) model of SRL: Task definition, Goal setting and planning, enactment, and adaptation. Zimmerman's (1990) model has three components: the use of self-regulated learning strategies, responsiveness to self-oriented feedback, and their interdependent motivational processes. Pintrich & De Groot (1990) outlined three components of self-regulated learning: planning monitoring and modifying cognition; students' management of their effort; and cognitive strategies used by students to learn, remember, and understand the material. While these three foundational models have been examined understanding the importance of social emotional context (Hadwin et al., 2007; Perry & Winne, 2006; Pintrich & De Groot, 1990; Zimmerman, Heart, Mellins, & Zimmerman, 1989; Zimmerman & Martinez-pons, 1990), the concept of self and socially regulated learning attempts to integrate the social dimension into the model of SRL (Järvelä, Järvenoja, Malmberg, & Hadwin, 2013). SSRL builds on the notion of self-regulation by considering co-regulation and shared regulation as additional strategies used with self-regulation. Co-regulation would represent two group members working on their own tasks that contribute to a group assignment while shared regulation would represent two group members working on a shared task. In the context of shared regulation the group weighs and negotiates multiple ideas and perspectives through metacommunicative awareness (Järvelä & Hadwin, 2013).

When considering the perspectives of others the concept of social perspective taking outlines that this can be helpful to broaden an individual's perspective assuming that there is attention paid to the interpretation and communication that leads to understanding dispositions of others along dimensions including emotions (Roan et al., 2009). This aligns with the belief that emotions may be the missing key to self-regulated learning (Op ' , Eynde, De Corte, & Verschaffel, 2007). In social contexts people need to clarify their emotion expression in order to support the social regulation of emotion (Reeck, Ames, & Ochsner, 2016). This is likely true because emotion regulation is a concept in emotional intelligence that is built on the foundation of perception of emotion meaning effective regulation can only occur if there is effective perception of emotion

(Mayer & Salovey, 1997). While the emphasis of this PhD thesis is on the measurement of Valid Self-Report of Emotion, the CPM conceptualization of the role of regulation of expression is that it is required for Valid Self-Report of Emotion as it occurs at the intersection of emotional communication, conscious regulation and unconscious regulation (see Figure 2.1.1). Given the necessity of regulation for the aim of measuring emotional communication this centralizes the need for effective perception of communication.

While section 2.1 described a range of conceptual approaches to emotion, and how to measure these emotions using SA and self-report instruments like PANAS, BEQ, and MES, in the second part of this literature review I specifically focus on how learning design decisions (e.g., emotion awareness tools, emotional sentence starters) made by teachers can influence emotions. When considering how to support emotional communication it is important to consider the intersection of learning design and emotion. For example, when students enter into interpersonal conflict a common coping strategy is to avoid dealing with the conflict by refocusing attention on the task (Näykki, Järvelä, Kirschner, & Järvenoja, 2014). Given this natural tendency, it may be useful to put the emotional focus on the learning material before a conflict arises. This strategy may help prevent interpersonal conflict. Potential learning design interventions like emotional awareness tools and emotional sentence starters may help students to effectively regulate and support their emotions when working individually and together with other peers in online tasks.

2.2.2 LEARNING DESIGN AND EMOTION

2.2.2.1 COMPUTER SUPPORTED COLLABORATIVE LEARNING (CSCL)

In Computer Supported Collaborative Learning (CSCL) so-called emotion awareness tools have demonstrated a variety of benefits to learning. Tools that focus on increasing the emotional awareness between collaborators have been shown to increase engagement (Arguedas, Daradoumis, & Xhafa, 2016; Järvelä & Hadwin, 2013); increase collaboration (Daradoumis, 2013); increase self-regulation (Arguedas et al., 2016); improve teachers attitude and feedback (Arguedas et al., 2016); increase social

support and interaction (Daradoumis, 2013; Feidakis, Daradoumis, Caballé, & Conesa, 2014); increase positive emotion after collaboration (Molinari, Chanel, Bétrancourt, Pun, & Bozelle, 2016); and increase transactivity (Molinari et al., 2016); Transactivity is an important concept in CSCL. “Transactivity indicates to what extent learners build on, relate to, and refer to what their learning partners have said or written during the interaction” (Noroozi, Teasley, Biemans, Weinberger, & Mulder, 2013).

There is also relationship between emotion awareness and learning outcomes as it has a positive correlation in some studies (Arguedas et al., 2016; Molinari et al., 2016). This intersection between awareness and learning supports the need to support student reflection and doing so is referred to in the CSCL literature as mirror supports which reflect data back to students about individual and group interactions (Järvelä & Hadwin, 2013).

2.2.2.2 UNIVERSAL DESIGN FOR LEARNING (UDL)

Universal Design for Learning (UDL) is the design perspective that learning environments should be designed up front considering the variability of students and provide flexible environments that anticipates different needs of students (Meyer, Rose, & Gordon, 2014, pg. 10). From the UDL guidelines according to checkpoint 5.2 I should provide students with multiple tools for construction and composition including so-called sentence starters (CAST, 2018). Sentence starters provide students with a menu of phrases to begin their communication in online discussions (Nussbaum, Hartley, Sinatra, Reynolds, & Bendixen, 2005). From a UDL perspective sentence starters are a tool that can be useful when considering how to support students in communicating their emotions by starting students to think about how they feel and then using sentence frames to communicate their emotional response (Posey, 2018).

2.2.2.3 SCRIPTING SUPPORTS INFORMED BY CSCL AND UDL

CSCL and UDL overlap in considering how to support students in online discussion using scripting to support written communication. While UDL provides a theoretical reason for supporting emotional communication CSCL provides empirical evidence for

the use of sentence starters. A vast body of literature has focused on managing complex group dynamics by using scaffolding or scripting. Scripts in group work have been used to help equalize participation in group communication, with some researchers indicating that some types of scripts (e.g., focused on group and social processes) had more influence on individual knowledge gain (Weinberger, Ertl, Fischer, & Mandl, 2005). One danger of scaffolding or scripting pointed out is to not get overly prescriptive (Dillenbourg, 2002; Rienties et al., 2012). Previous research has highlighted that individual differences and personal traits of students significantly influence attitudes and behaviour in contributing to discourse (Knight et al., 2017; Mittelmeier, Rienties, Tempelaar, & Whitelock, 2017; Rienties et al., 2012) and group processes as a whole, so scaffolding needs to take into consideration that students might react differently to scaffolding. The use of sentence starters to augment writing have been described as most appropriate for college students with applications to other contexts (Newell, Beach, Smith, & VanDerHeide, 2011).

I will distinguish between cognitive and emotional scaffolding. The cognitive script that has more of a focus on the argumentation and its relationship to datum (Weinberger, Stegmann, Fischer, & Mandl, 1997). Several researchers have found positive evidence of providing cognitive sentence starters and scaffolds to encourage participation (Belland, Walker, Kim, & Lefler, 2016) as indicated in a meta-review of 144 experimental studies which found that computer-based scaffolding and in particular cognitive scaffolds had a consistently positive impact on cognitive outcomes. An alternative form of scaffolding is to ask participants about their emotional reactions to the learning material (i.e., emotional sentence framing). By introducing emotional sentence starters, I anticipate that this may remove a barrier to participation for reluctant students. If students make comments that explicitly state their emotional reactions to learning material, it may actually help them to even the playing field of interpretation, given that emotional inference from text has been linked to the capacity of working memory of the reader (Gillioz, Gygax, & Tapiero, 2012). In CSCL it has been proposed that sentence starters might be useful when explicitly supporting the phases of self and socially regulated learning. These phases include the phase of reflection (Järvelä & Hadwin, 2013). From this perspective emotional sentence framing has the potential to

improve inference from text while in use as well as support reflection on the task after using the scaffold connecting ideas of CSCL and UDL.

2.3 CONCLUSION

When pursuing the development and evaluation of sentiment analysis for the proposed use as Emotion Learning Analytics (ELA) this review has indicated that there are implications in terms of 1) how to evaluate ground truth based on the emotional theory selected by researchers, and 2) design decisions in learning environments that may be beneficial in terms of supporting emotional communication. This review has highlighted how the CTE may be an appropriate theory for the basis of sentiment analysis in the context of learning to create ELA. The CTE implies that considering student perception on the emotional content of communication is critical to determine the extent to which an ELA measure using sentiment analysis aligns with the collective intentionality of the students. At the same time the review of existing use of sentiment analysis in the context of learning demonstrated limited exploration of accuracy in terms of the CTE. Furthermore, there is evidence that students have the capacity to identify valence in text. It is because this approach is both theoretically founded and there is an identifiable gap that I raise research questions focused on training a sentiment analysis classifier based on student perceptions. Which raises research question 1 (RQ1): To what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters? RQ1A: To what extent do students agree in terms of inter-rater agreement when providing examples? RQ1B: To what extent do Mechanical Turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples?

When reviewing theoretical models of valence, which is commonly measured using sentiment analysis, I explore multiple perspectives. By going in depth on Mosier's (1941) study I demonstrated how the interpretation of text may require considering that valence may be more complex than the bipolar perspective which considers valence comprised of positive, negative, and neutral. The Evaluative Space Model (ESM) expands the bipolar dimension of valence (positive, negative, and neutral) to a plane considering four possible valence categories: positive, negative, neutral, and mixed.

When reviewing 15 existing sentiment analysis studies in the context of learning I found that less attention has been paid to the categories of neutral and mixed which was consistent with the claim from the one study which measured all four categories and mentioned that not enough attention has been paid to the categories of neutral and mixed valence. Which raises research question 2 (RQ2): To what extent can crowd sourced, and in particular student sourced, examples train a machine learning classifier to predict the valence categories of positive, negative, neutral, and mixed? RQ2A: To what extent can student labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2B: To what extent can Mechanical Turk labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2C: How do logistics classifiers trained using student labels and Mechanical Turk labels compare to general benchmarks when predicting the valence categories of positive, negative, neutral, and mixed? RQ2D: To what extent do students find predictions from a student sourced classifier useful?

In summary, there is a theoretical reason to explore how student perceptions of emotion in text relates to sentiment analysis and an associated gap in the research. There is also a theoretical reason to consider a valence model that considers text to be in one of four categories (i.e., positive, negative, neutral, and mixed) and an associated gap. It is for these reasons that I propose creating a classifier based on student perceptions that considers valence as four categories. There is also reason to consider how the design of the learning environment influences the measurement of emotion. I next explore how research methods can use design as part of research methods to explore emotional measurement.

To examine the relationship between learning design and learning measurement I explore how emotional sentence starters might support students to identify emotion expression by using DBR and experimental design (see section 3.3.1). While the literature review provided a theoretical reason to consider emotional scripting DBR in conjunction with experiment design provides a method to examine research question 3 (RQ3): To what extent can emotional sentence starters improve the inter-rater reliability of student examples?

In terms of considering how reliable the student sourcing method is to generate a sentiment analysis classifier I apply the method on each of the two new datasets through the same process to train a classifier which raises research question 4 (RQ4): To what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?

In reviewing psychometric instruments, I consider how to collect data about state and trait of emotion to examine how those measures align with a student sourced sentiment analysis classifier. The reason to determine the extent to which measure of emotional state and trait correlate with measures of emotional expression is to identify the extent to which a measure of SA relates valid self-report of emotion as defined by the CPM theory which raises research question 5 (RQ5): To what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?

As the student sourced labelling approach is aligned with the CTE and our review indicated that SA studies in the context of learning use a variety of theories on emotion I contrast the examination of correlates with a classifier trained on student sourced labels with a SA technology built for general purposes which raises research question 6 (RQ6): To what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?

CHAPTER 3 METHODOLOGY

3.1 INTRODUCTION

In Chapter 2 I examined the existing literature to identify gaps regarding emotional measurement in the context of learning. In outlining the literature on measurement, I illustrated the relationship between design and measurement, which raised a range of research questions about both how to measure emotion in the context of learning as well as how design can influence the measurement of emotion in online learning. This review led to six specific research questions about measurement, design, and the interaction between design and measurement. The literature review also navigated through a multi-layered debate about the nature of emotion to establish a proposed model, the univariate mixed emotion (UME), for the purpose of having a theoretical basis for measurement. The theoretical UME model of emotion provided attributes, such as the dimension of valence, organized as both bipolar and bivariate using the UME paradigm. Furthermore, I established a position that students themselves may be judges in terms of categorizing their messages as positive, negative, neutral or mixed. To investigate our six research questions associated with evaluating a measure in terms of validity and reliability it is necessary to formalize the theoretical perspective for the ontology, epistemology, and methodology to undertake such an investigation.

3.2 ONTOLOGY AND EPISTEMOLOGY

When establishing a theoretical framework for research it is important to consider how ontology (nature of being) and epistemology (theory of knowledge) informs methodology (Cohen, Manion, & Morrison, 2007). The focus of this thesis is on the measurement of emotion, from the perspective of the CTE, where emotions are a collective intentionality (Barrett, 2012). This is a claim that is post-positivistic, indicating that there is no objective reality. Barrett argues that while there is no objective reality there is collective intentionality illustrated with the example of how people consider the distinction between a weed and a flower. To connect this position to the Open University I illustrate how collective intentionality has interpreted the plant

called ragwort. After discussing the identification of ragwort as a weed, I provide some personal insights, on how text communication is informed by collective intentionality, to articulate my ontological and epistemological view on emotion expression in text.

In a given context, such as Milton Keynes, there is a collective intentionality to remove ragwort as it is considered a toxic weed and can kill grazing animals. There is in fact a Weeds Act of 1959 in the UK which aims to prevent the spread of weeds, including ragwort. Failure to prevent ragwort from spreading on your land could result in a one thousand pound fine. Because of the collective intentionality to have grazing animals in areas of the UK, including Milton Keynes, the plant is viewed as a weed. There is little question that Ragwort is a weed because of how these properties lead to legislation that it should be destroyed. However, ragwort was rated in the top ten in nectar production for UK plants. It is plausible that if there were more bee keepers than cattle farmers that the plant would be referred to as a flower and intentionally planted as a means of nectar production. However, when walking through the Ouzal Valley Park towards the Open University it is immediately obvious that there are many farmers with grazing animals. Students at the Open University would likely understand, given the properties of ragwort (i.e., toxic to grazing animals and producer of nectar) that in the context of Milton Keynes ragwort would be considered a weed rather than a flower.

When asking if it is real to call ragwort a weed there is clear evidence (in the form of legislation) of collective intentionality that ragwort is really a weed. Barrett argues that emotions are real in the same way that weeds are real. Given the position that what makes emotions real is a collective intentionality. Barrett argues that when considering accuracy for the detection of emotion it is best to consider consensus. While a strict viewpoint on the relativistic nature of reality would suggest there is no quantifiable way of establishing knowledge, collective intentionality suggests that consensus by members of a social context is the most viable approach toward considering the accuracy of detecting emotion. To connect this to the viewpoint that students may be the best judge of emotion expression in text from their group discussion, this would parallel the example of ragwort by saying that students at the Open University would have sufficient knowledge to make the distinction between classifying ragwort as a weed verse flower.

Perhaps students from another university situated in a place where beekeeping was a dominant practice might classify ragwort as a flower.

To summarize the ontological view of emotions, emotions are real in terms of collective intentionality. To have knowledge of emotions expressed in text requires perception informed by collective intentionality, such as a person from the context where the text expression was authored. For the context of online group work collective intentionality suggests that students themselves are likely to be the best judges of emotion expression in their own text messages. When considering how accurate students are at judging emotion expressed in text consensus is the suggested approach as it aligns with the perspective of collective intentionality.

From a personal perspective I have lived in a variety of contexts including: locations in the United States such as the Lummi reservation, Maui Hawaii, Seattle Washington, and Boston Massachusetts; Also, briefly in Milton Keynes in the United Kingdom. Each context uses emotional expression in communication differently. In the Lummi reservation people often use the coastal Salish phrase of “Huy ch q’u Siam” to say what loosely translates into “thank you respected one”. This phrase is such a common honorific that anyone in the community would readily identify it as an acknowledgement of respect. I have used this phrase in chat discussions with family that live on Lummi. The last time I was on Lummi after dropping off a hitch hiker they said “Huy ch q’u” (i.e., thank you). Similarly, in the context of Maui a phrase that has contextual meaning is “broke da mout” which loosely translates into “the food is so good that it will break your mouth”. When using text to communicate with my friends from Maui this phrase is used when describing a nice restaurant that has recently opened. In Seattle the term “Ave rat” refers to homeless youth on University avenue (a street near the University of Washington). It is a phrase a college student might refer to themselves as in a joking manner if they spend a lot of time on University avenue (e.g., “I feel like such an ave rat these days”). In Boston, a common expression is “Wicked Smaaht” which can be used as a compliment indicating someone has knowledge. It can also be used as a criticism when paired with a criticism of intelligence (e.g., “he’s so wicked smaaht he can’t tie his shoes!”). In Milton Keynes, it took me some time to understand what someone meant when they said something was “pants”. For example,

in online text messages with friends in the UK they have used the phrase “this event is pants”. In the US pants refers to trousers while in the UK pants refers to underwear. To call something pants in the UK is a manner of expressing that it is not good. I am still not entirely sure why residents of the UK have a negative association with underwear.

All of these examples from Lummi, Maui, Seattle, Boston, and Milton Keynes demonstrate how a collective intentionality shapes how I interpret text communications. Every place I have lived required insider perspective to interpret communications. In each of these contexts I have to some extent self-identified as an outsider. In fact, in Hawaii the phrase “wat tryin’” is something a person would say if they thought someone else, who was not native to Hawaii, was using colloquial expressions to fit in. In this PhD thesis I am not trying to fit into the context of research (a university in the Netherlands). Instead, I rely on people from the context to interpret the emotional meaning of text communication as they are part of a collective intentionality. In adopting the perspective of collective intentionality, I am acknowledging my position as an outsider of the collective and defer to people in the context for insight into how emotion is expressed.

To connect this perspective to an example communication from a one of the Studies in this PhD thesis, I suspect the Dutch have an affinity with cheese. My suspicion is based on my limited exposure to the context. When living in Seattle my standard cheese purchase included a gouda cheese from the Netherlands called Olde Amsterdam, indicating cheese is a global Dutch export. When visiting the Netherlands cheese was one of the main items for sale at the Saturday market. Just as I saw grazing animals as I walked to the OU students in the Netherlands would see cheese in the market when walking to their university. When packing up the remaining groceries at the end of writing camp in the UK my Dutch supervisor brought a left-over assortment of cheeses he had in his cabin. With this minimal contextual experience of associations between cheese and the Dutch I believe this helps illustrate how collective intentionality of students in the context of a University in the Netherlands identified the following comment as positive: “The Netherlands are pretty good, go for the cheese guys”. While not all students in this context are from the Netherlands I suspect they would work toward a collective intentionality just as I did when I was in the UK.

The aim of this work is not to define how communication at a University in the Netherlands is contextually different, but rather to rely on the perspective of students in the context to be experts to generate a classifier that is sensitive to the context. Presumably, if the process of student labelling can generate a classifier that is context sensitive then it may be a process that has application to a variety of contexts such as Lummi, Boston, Seattle, and Milton Keynes. This perspective represents the ontological view that emotions are real via collective intentionality, and epistemologically speaking I know emotional expression in text through contextual understanding. While this viewpoint more generally aligns with relativistic perspectives Barrett's suggestion that consensus is a way to consider accuracy of emotion expression connects a relativistic point of view to a quantitative approach by considering how I might model consensus of student perception. To model consensus in a university setting I have selected a positivistic approach using quantitative methods.

3.3 OVERVIEW OF ADOPTED APPROACH

3.3.1 DESIGN BASED RESEARCH AND EXPERIMENTAL DESIGN

This thesis uses a combination of two main learning science principles, namely Design Based Research (DBR) and experimental design. The reason why I have chosen to use a DBR approach is to explore how to support students to reflect on emotional communication from their own group discussions. Effectively the aim of this PhD thesis is to explore the extent to which I can model consensus about what emotion is expressed in text based on student perceptions. As I anticipate that there is a theoretical limit to the extent to which they will agree I explore how design can support that perception. The aim of this PhD thesis is the creation SA which I refer to as Emotional Learning Analytics (ELA). In the field of Learning Analytics DBR is aligned with the observation that there is a dependency between learning design and learning analytics.

One obvious criticism of DBR is that it is typically iterative, and often a combination of quantitative and qualitative approaches. To counteract some of these criticisms, as I designed one experiment with a follow-up experiment over a period of two years, I opted to specifically use an experimental design approach for the second

study. When considering multiple research conditions the “gold standard” is to have a randomized control trial (Alana & Snibbe, 2006). When designing a randomized control trial, parallel design has two distinct groups where research subjects are assigned one condition throughout the entire study (Siepmann et al., 2016). With parallel design randomization is a powerful tool to ensure validity (Siepmann et al., 2016). As the parallel design has the strongest benefit from randomization in terms of validity and the intention the parallel design has been selected for our second experiment. In addition, when conducting a randomized control trial, the results can generate causal conclusions (Hutchison & Styles, 2010). As the experimental condition is planning to use emotional design to support emotional communication the results will have the potential to demonstrate the causal effect of emotional design to support emotional expression should the design cause an increase in accuracy for emotion detection.

3.3.2 THREE STUDIES TO TEST THE SIX RESEARCH QUESTIONS

The structure of Chapters 4-6 is organized by three studies based on the results of two Experiments. All experiments were conducted in the same business school in two consecutive years in the Netherlands. The respective university recruited international business students as part of the teaching philosophy at the University was that students could learn from a diverse group of peers. In this context students typically had a problem-based learning (PBL) curriculum, meaning that they were used to working in groups to solve a specific problem, see Tempelaar et al. (2015; 2017) for a detailed description of the educational context. As described in greater detail in Mittelmeier et al. (2018), students worked together on an online task in a computer lab. The online task consisted of a World-Bank assignment, which asked students to discuss data from a set of countries and work on a problem of making a funding decision in a group.

Students were typically assigned randomly to groups of 5 ($M=4.73$ $SD=0.84$) in a laboratory setting, whereby each student had a desktop computer, and all written communication was online as part of a regularly occurring lab session for their course. The discussions took place in Udio, which is a platform designed to support reading comprehension through providing short high interest content with integrated reading comprehension supports, such as discussing the reading (Hillaire et al., 2018). Although

the details of the two experiments were slightly different, overall the two experiments followed broadly the following structure illustrated in Figure 3.3.2.

Input	Process			Output
BEQ (5 min)	Ice Breaker (5 min)	Discuss World Bank Data (50 min)	Provide an Answer (5 min)	Post Activity Label Messages (60 min)
Before Lab	During Lab			After Lab

Figure 3.3.2 Study design of Experimental Study 1 and Study 2

In Experiment 1 I conduct Study 1 which is a SSSAC, whereby I report the detailed narratives in Chapter 4. This is effectively a quantitative study that uses crowd sourcing methods to establish labels that represent the social reality of what valence is perceived in text communications by students from the context. The result of the crowd sourced labels is used to train a machine learning classifier which learning from the social perspective of the students in the context. The accuracy of the classifier is checked using cross validation on the labelled data and the overall accuracy of the classifier as well as the accuracy for each univariate mixed emotion valence category is compared. The benchmark technologies are bivariate measures which I interpret in a mixed bivariate manner where mixed is an inferred value. By comparing the univariate mixed emotion classifier with the mixed bivariate measures, I am testing to see through a quantitative method if the UME model out-performs bivariate mixed emotions models. This positivistic quantitative study attempts to detect social norms of the detection of valence in text is a methodologically straight forward examination that uses best practices associated with evaluation of machine learning classifiers and operates on the premise that the social perspective has a normative paradigm that can be evaluated quantitatively.

From this stable starting point the follow up Experiment 2 is conducted using a randomized control trial, such that the control condition is a replication of the first experiment, and the intervention investigates the influence of DBR considering how emotional sentence starters influence the study. The design being tested is emotional design to support the perception of emotion in text expression as I evaluate how the use

of sentence starters changes the generation of a student sourced sentiment analysis classifier (SSSAC). The design decision is informed by the UDL guidelines making this intervention the aspect of this study that is informed by neuroscience through the translational framework of UDL. This intervention is there for the most methodologically questionable as there are criticisms that neuroscience to education is a bridge to far and none of researchers involved in the study have a neuroscience background. However, one of the researchers does have a background in UDL and the platform used in the study was developed as a UDL platform with integrated UDL supports. The methodologically questionable elements of DBR and Educational Neuroscience are also supported by having a sound design for Study 1 and using the control condition for a replication in Study 2. For the detailed narratives, I refer to Chapter 5.

Finally, Study 3 is an attempt to use triangulation of emotional measurement to better understand the influences of the emotional sentence starters on the activity. The input process output model is used to anchor emotional measures at the input designed to measure student dispositions, self-report of emotion integrated into the beginning of the activity to test the influence of incidental emotions on the activity and finally using multiple validated psychometric surveys as exit questionnaires after the activity to get a measure of emotion at the output of the activity. This battery of triangulation is the second measure taken to ensure that the DBR intervention of emotional design is rigorously examined. The details of Study 3 can be found in Chapter 6.

Using Study 1 and Study 2 to examine accuracy for SSSAC I establish a foundation to interpret external validity in Study 3 which focuses on correlates with psychological measures of emotion. By considering both internal and external validity this design examines both the accuracy of a classifier based on student perceptions of emotion expressed in text (Study 1 and Study 2) and the accuracy of the classifier in terms of its relationship to the emotional experience of students (Study 3). In a schematic overview Table 3.3.2, I have summarized the main characteristics of the two Experiments, and the three respective studies.

Table 3.3.2 Overview of Study 1, Study 2, and Study 3

	Study 1	Study 2	Study 3
Experiment 1 2016 (N=767)	Train and Benchmark Classifier	Test effects of Emotional Sentence Starter on reliability of student labels (RCT). Optimize Classifier Considering Reliability, amount of training data, pre-processing, and processing	Correlation analysis between psychological measures in experiment 2 with predictions on valence of messages in experiment 2 (using classifier trained in experiment 1)
Experiment 2 2017 RCT (N=884)	Control (n=437)	Test Classifier and Interview Students	
	Intervention (n=447)		

3.4 INSTRUMENTS USED

As indicated from the literature review Chapter 2, there are many ways to measure emotion. In this section I briefly provide an overview of the measures that have been collected by students, which will be discussed in the order of Figure 3.1. Note that not in all three studies all these instruments were used. For the specific details of which instruments were used for which study, I refer to the respective Chapters 4-6.

3.4.1 PRE-QUESTIONNAIRES

3.4.1.1 BERKLEY EXPRESSIVITY QUESTIONNAIRE (BEQ).

Prior to participating in the study students filled out the Berkley Expressivity Questionnaire (BEQ). The questionnaire has three constructs: Negative Expressivity, Positive Expressivity, and Impulse Strength. Negative expressivity aims to measure the extent to which students are comfortable with expressing negative emotion. Positive expressivity aims to measure the extent to which students are comfortable expressing positive emotion. Impulse Strength aims to measure the extent to which students are capable of regulating their emotion expression. Each is measures on a scale from 1-7. The students were given the questionnaire prior to the study with four questions from each construct for a total of 12 items. The entire tool is 16 items. Four items were not administered after consulting the teacher at the site for appropriateness of the questions for the student population. The entire tool and the items administered are in appendix 1.

3.4.2 DATA COLLECTED DURING THE COMPUTER LAB

3.4.2.1 UDIO

To conduct the two experiments I used Udio which is a platform that provides options for engaging with reading material, including embedded discussion with peers, and interactive features where readers can provide reactions to the reading (S. A. Crossley & McNamara, 2016). When supporting discussions Udio provides the option to include sentence starters, which are provided following the UDL literacy by design e-book features (Coyne, Pisha, Dalton, Zeph, & Smith, 2012). Udio also allows readers to provide a reaction to reading by selecting between one and twelve emotional words to provide a reaction with the feature React (Hillaire et al., 2016). For this PhD thesis Udio provided a case study where students were asked to make a funding decision by examining the educational achievement data from their home countries which has previously been used in conjunction with Udio and demonstrated increased engagement in discussion (Mittelmeier, Rienties, Tempelaar, Hillaire, & Whitelock, 2018).

3.4.2.1.1 REACT

React was developed using Design Based Research (Hillaire et al., 2016) to seek a universal self-report mechanism for students to provide reactions to their readings. Figure 3.4.2.1.1 visually illustrates how students could self-report their emotional reactions using a multi-select of twelve emotion words. React provided the following list of words as options: engaging, interesting, challenging, curious, calming, good, dull, boring, sad, confusing, frustrating, and annoying. In Figure 3.4.2.1.1 the words engaging, interesting, challenging, good, boring, and sad have been selected as an example of what the student would see after selecting. Previous work has examined how the use of React relates to sentiment analysis of discussions in Udio (Hillaire, 2015; Hillaire et al., 2018). It is in fact this previous work that inspired the topic of this PhD thesis.

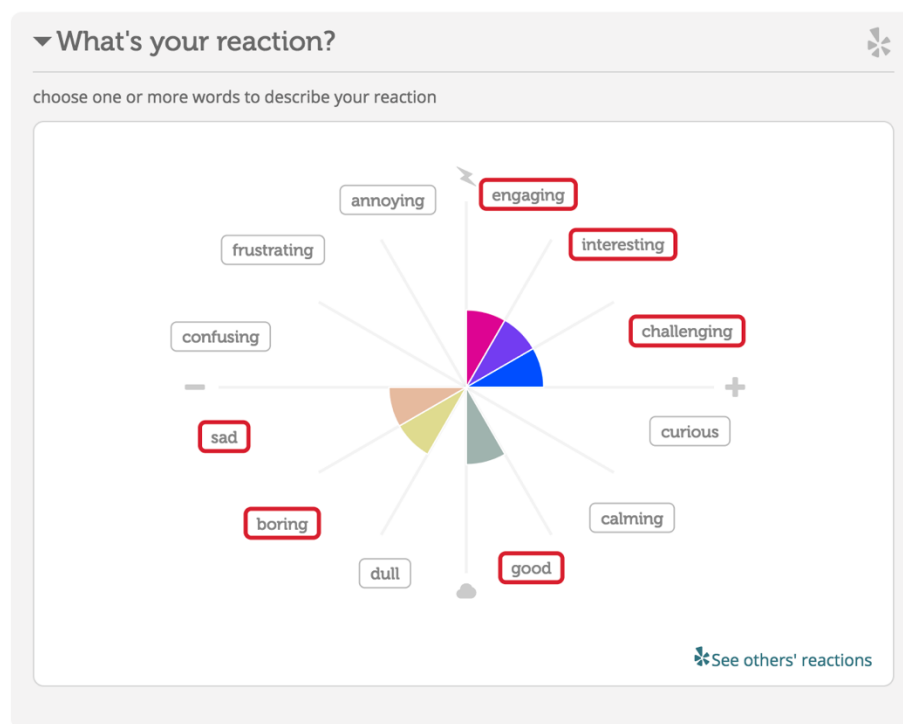


Figure 3.4.2.1.1 Interface of 'React' to self-report emotional response

3.4.2.1.2 DISCUSS IT

When reading text in Udio students have the ability to discuss the reading with their peers using Discuss It. Discuss it offers a chat window where students can comment on the reading (see Figure 3.4.2.1.2). In the experiments of this thesis students are using Discuss It which is the same interface used for previous work (Mittelmeier et al., 2018) where students talk about a case study where they are working as a group to make a funding decision. The interface provides a synchronous discussion.

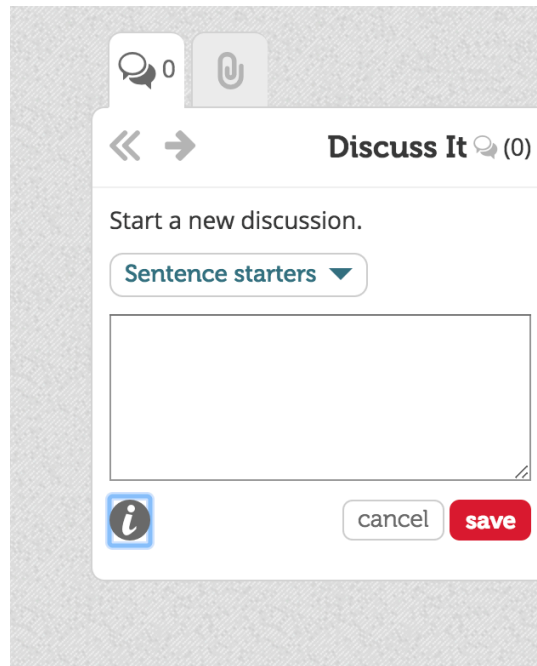


Figure 3.4.2.1.2 Discussion interface with Sentence Starters

3.4.2.1.3 ITERATING THE DESIGN OF UDIO WITH EMOTIONAL SENTENCE STARTERS

Udio is a platform created using design based research methods (S. A. Crossley & McNamara, 2016). For example, the development of self-report instrument called React occurred through a series of iterations (Hillaire et al., 2016). When I was working as an educational software architect on Udio, I was very curious about the intersection between providing a reaction to a reading and what emotional communication is occurring in the discussion. This curiosity lead to some preliminary analysis which

indicated there were some relationships between SA measures of discussions and reactions provided by use of React (Hillaire, 2015). It was from participating in the design work which raised a question about the potential to use the sentence starter feature to support emotional expression.

To iterate on previous work, in this PhD thesis I introduce emotional sentence starters (ESS) where the sentence starter feature of Udio is used to provide ESS by asking students to start Discuss It comments using sentence starters “I had a [positive | negative | neutral | mixed] reaction to...”. The intent of this design is to have students start their comments by contextualizing what they are about to say with the valence that categorizes their reaction to reading material.

3.4.3 DATA COLLECTED AT THE END OF THE COMPUTER LAB

Once students completed using Udio they were asked to respond to an exit survey which included PANAS and MES. The purpose for including these two measures is to follow the advice that emotional measures should model the entirety of the construct (Weidman et al., 2016). In this case the Evaluative Space Model (ESM) considers valence comprised of positive, negative, neutral, and mixed experiences (Cacioppo et al., 1999, 2004).

3.4.3.1 PANAS

PANAS is commonly used in emotion research (Drake, Myers, & Drake, 2006) and it is the instrument that best represent bivariate perspectives on valence (Green et al., 1993; Leue & Beauducel, 2011; Watson et al., 1988, 1999). According to google scholar PANAS has been cited over 32,000 times. PANAS produces largely independent scores for positive and negative affect (Feldman Barrett & Russell, 1998). However, PANAS was developed considering the parallel perspective on valence (Watson et al., 1988). The Evaluative Space Model (ESM) proposed by Cacioppo et al. (1999) points out that bivariate measures are sufficient to measure emotions in some circumstances depending on where participants are in the evaluative process. However, bipolar (integrating positive and negative) approaches can at times do a better job of measuring emotion

(Cacioppo et al., 1999; Feldman Barrett & Russell, 1998). The entire tool was administered and is detailed in appendix 2.

3.4.3.2 MIXED EMOTION SCALE (MES)

The Mixed Emotion Scale (MES) takes a distinctly different approach from PANAS towards measuring valence from both a theoretical perspective and from a practical perspective. From a theoretical perspective the MES aims at explicitly measuring the integrative nature of positive and negative valence and was shown to be distinguished as a measure from the related measures (Berrios & Totterdell, 2013) of ambivalence (Pekrun et al., 2011) and intolerance of ambiguity. The MES does this from a practical perspective by asking participants to rate their agreement with statements that describe both parallel and integrative experiences. Rather than asking participants to describe positive and negative aspects of their experience using ratings for emotion words (like the PANAS does), the MES asks participants to rate their agreement with statements that describe parallel and integrative emotional experiences. For example, the MES asks participants to rate their agreement with the statement “I felt a mixture of emotions” (Berrios & Totterdell, 2013). By directly measuring the experience of mixed emotions the MES is best described as a measure based on the univariate mixed paradigm (Kreibig & Gross, 2017). The entire tool and the items administered are in appendix 3.

3.4.4 DATA COLLECTED AFTER THE COMPUTER LAB

3.4.4.1 POST-SURVEY

In the post activity we had students label the sentiment of messages from their own group discussions. As the literature review on rating messages suggested 1) fatigue can reduce rating quality 2) when rater quality is high it is better to get single ratings on more messages over than to get multiple ratings on fewer messages, 3) from the CTE perspective we consider people from the social context as experts, and 4) the aim of SA is to model the opinion of the author and the reaction it elicits by the intended audience, which in our case is students, we consider student opinion the goal of the measure.

Based on this premise we ask students to review their own group discussions and first estimate if there are any messages in each valence category and follow that estimate with providing 1-3 examples of messages for the valence category (See appendix 4 & 5 for complete instructions). As students are providing examples the method generates many single ratings and some ratings with overlap.

3.4.4.2 SEMI-STRUCTURED INTERVIEWS

There were three phases to the semi-structured interviews: 1) rating messages, 2) comparing ratings with predictions from the algorithm, and 3) answering open ended questions.

First, students examined a random subset of messages from their own group discussions. Students were first asked to rate the messages as positive, negative, neutral, or mixed.

Examine Comments

Comment	Valence	Notes
2027) Hi, my name is [Name] and I am from [Country].	<input type="checkbox"/> Positive <input type="checkbox"/> Negative <input checked="" type="checkbox"/> Neutral <input type="checkbox"/> Mixed	
3023) I think a country such as Lebanon would be more deserving as in a global comparison they are rock bottom	<input type="checkbox"/> Positive <input type="checkbox"/> Negative <input checked="" type="checkbox"/> Neutral <input type="checkbox"/> Mixed	

Figure 3.4.4.1A Interview Protocol – Examine Comments

Second, students reviewed the same messages again. This time the students saw 1) the message text, 2) the text features the SSSAC Logistic used to make a prediction, and 3) the prediction made by SSSAC Logistic. Finally, they were instructed to identify if the prediction agreed with the label from the first section (yes/no). In the event of disagreement, they were asked to examine the features used to make the prediction and identify if the prediction and the text features changed their mind (yes/no).

Sentiment Analysis Details

Comment Analysis Details	Valence	Agreement	Change
2027) Hi, my name is [Name] and I am from [Country].['.', ',', ('and', 'i'), 'hi', ('.', 'my', 'name'), ('my', 'name', 'is'), ('.', 'my'), ('my', 'name'), ('name', 'is'), 'name', ('i', 'am'), ('i', 'am', 'from'), ('am', 'from'), ('and', 'i', 'am')]	Neutral	[x]Yes [.]No	[.]Yes [.]No
3023) I think a country such as Lebanon would be more deserving as in a global comparison they are rock bottom[('i', 'think'), 'think', ('would', 'be'), 'would', ('in', 'a'), 'comparison', ('be', 'more'), 'country', ('a', 'country'), ('they', 'are'), 'lebanon']	Mixed	[.]Yes [x]No	[x]Yes [.]No

Figure 3.4.4.1B Interview Protocol Sentiment Analysis Detail

The third part of the interview asked if the students found the classifier useful. The students wrote an open-ended text response to answer the question: “Do you think sentiment analysis of your comments is useful when provided for individual messages?”

Q&A about the utility of sentiment analysis

Question	Answer
Do you think sentiment analysis of your comments is useful when provided for individual sentences?	I think it was useful in showing me an alternative argument. There were more than one case that changed my mind which shows that I learned from it.

Figure 3.4.4.1C Interview Protocol – open ended questions

3.4.5 DATA AND SENTIMENT ANALYSIS TOOLS

3.4.5.1 SENTISTRENGTH

SentiStrength is an example of a lexical approach (Thelwall, 2013), which builds a list of positive and negative words. After identifying the presence of positive and or negative words SentiStrength uses additional strategies to improve the quality of the prediction, including considering negations, boosting words, and emoticons. Negations words like ‘not’ are used to identify when the valence of a word is negated. For negations, SentiStrength reverses the valence of the words in the communication and reduces the magnitude of the valence score by half. In contrast, boosting words are thought to intensify the meaning of a word. For example, ‘very’ is interpreted to mean that the magnitude of the valence should be increased. SentiStrength also used Emoticons. Emoticons refer to constructing an image in text through the use of characters that looks like a facial expression (e.g., indicate a smiling face with “:-)“). SentiStrength has the options to make a bipolar prediction of positive, negative, or neutral as well as bivariate prediction of a separate score for both positive and negative valence in text communication (Cacioppo et al., 1999; Thelwall, 2013) which makes the technology theoretically aligned with this investigation that includes mixed emotion.

When evaluating SentiStrength, the published datasets of ground truth labels are comments from a variety of social media (i.e., YouTube, BBC, DIGG, Runners World Forum, MySpace, and Twitter). SentiStrength identified the occurrence of positive and negative words based on a lexicon of positive and negative words (Thelwall et al., 2010). For this PhD thesis I contacted the author of SentiStrength and got the latest distribution of the code at the time (SentiStrength_DataEnglishFeb2017) for our analysis. I use SentiStrength as one of the benchmarks in Study 1.

3.4.5.2 VADER

VADER is an example of a SA technology that used a heuristic approach (Hutto & Gilbert, 2014) meaning that heuristic rules capable of predicting the valence of text are used to create the technology. VADER is an example of a technology that claims to

have higher than human levels of accuracy (Hutto & Gilbert, 2014). The result of a VADER prediction is three probabilities of valence. A probability for Neutral, Positive, and Negative. While it does not directly interpret mixed emotion the predictions of positive and negative as two separate values is a bivariate representation of valence. I use VADER as one of the benchmarks in Study 1.

3.4.5.3 SÉANCE

Séance is a tool that includes multiple indices for SA (available at: <http://www.kristopherkyle.com/seance.html>). Séance provides a single interface to use a variety of freely available SA technologies. I also used the co-occurrence paradigm when interpreting the bivariate predictions from Séance. GALC, the Geneva Affect Label Coder, attempts to categorize 36 affective label experiences and is capable of providing a positive and negative score that reflects the detection of those affective experiences regarding valence (Scherer, 2005). GI, the General Inquirer, includes the Harvard IV-4 dictionary lists manually constructed containing over 11,000 words, and is the oldest list in widespread use (S. A. Crossley, Kyle, & McNamara, 2017). EmoLex is a crowd sourced SA technology which predicts discrete emotions as well as bivariate valence developed for general purposes (Mohammad & Turney, 2010, 2013). EmoLex also used part of the General Inquirer for its creation (Mohammad & Turney, 2013). Hu-Liu uses a list of over 2,000 negative words and over 4,000 positive words mined from product reviews (S. A. Crossley et al., 2017; Hu & Liu, 2004; B. Liu & Street, 2005). Lasswell is a lexicon of 63 word lists organized into nine categories (power, rectitude, respect, affection, wealth, wellbeing, enlightenment and skill) (S. A. Crossley et al., 2017). Lasswell is also included in the General Inquirer (Lawrence & Rodriguez, 2012).

For Study 1 I used Séance to generate bivariate predictions using GALC, GI, EmoLex, Hu-Liu, and Lasswell. While Séance is capable of making predictions based on VADER, for Study 1 I used a more recent version of VADER than the version included in Séance.

3.4.5.4 LIWC

The Linguistic Inquiry and Word Count (LIWC) is a dictionary approach to SA. The technology is based on dictionaries that are composed of almost 6,400 words, word stems, and selected emoticons. LIWC is a common benchmark technology which was even used to benchmark Séance (S. A. Crossley et al., 2017). I use LIWC as one of the general approaches as a benchmark in Study 1.

3.4.5.5 MAJORITY CLASS BASELINE (MCB)

The majority class baseline (MCB) always predicts the most common category of message based on the training set of data. In this study the training set of data is the examples provided by students. In this Study students provide more examples of positive messages than any other category so the MCB simply always predicts messages to be positive. MCB is a common heuristic algorithm used to benchmark accuracy to ensure that a measure achieves a higher level of accuracy than simply always guessing the most common answer when predicting a category for text messages. I use MCB as one of the benchmarks in Study 1.

3.4.5.6 RANDOM BASELINE (RB)

The random baseline (RB) simply randomly guesses a category for text messages. In this study I classify messages as positive, negative, neutral, and mixed so the RB randomly categorizes messages into one of these four categories. Given I have four categories, each category is predicted by the RB roughly 25% of the time. The reason to include RB as a benchmark is to evaluate which classifiers do better than randomly guessing. RB is used to benchmark SA classifiers to understand how they perform in comparison with random guessing. I use RB as one of the benchmarks in Study 1.

3.4.6 MACHINE LEARNING CLASSIFIERS

Shickel et al. (2016) recommended the detection of positive, negative, neutral, and mixed for valence detection and found that the Bag-Of-Words classifier performed best in comparison with six benchmarking technologies. It is for this reason I focus on using

a Bag of Words classifier. When reviewing AL studies of SA (see section 2.1.2) a study suggested that Naïve Bayes (NB), and Support Vector Machines (SVM) outperformed most algorithms for SA (Altrabsheh et al., 2013). It is for this reason I also train a NB and SVM classifier during the benchmark analysis of the proposed Logistic Regression.

3.4.6.1 PRE-PROCESSING TEXT

When generating a SSSAC, students operate as a crowd to evaluate the emotional content of messages from their own group discussions. I used student sourced labels to train the classifier (see section 3.5.1). The examples provided by students were used to evaluate a classifier and/or to train a classifier. The examples were then processed to determine what features to use from the text.

When determining features for a text classifier, it is common to select a threshold value n , and extract text features that are below n for n -gram parsing, where n -grams represent co-location of words that are n in length. For example, the phrase “This is a good point” has five monograms {“this”, “is”, “a”, “good”, “point”}; four bi-grams {“this is”, “is a”, “a good”, “good point”}; and three trigrams {“This is a”, “is a good”, “a good point”}. When parsing text frequently stop-words (commonly used words) are removed prior to computing features (e.g., n -grams) from text (Lonchamp, 2012; Wiebe, Wilson, & Cardie, 2005). In the example “This is a good point” when removing stop-words the phrase becomes “good point” as the first three words are commonly occurring words. The phrase that remains after stop-word removal, “good point”, has two monograms {“good”, “point”}; and one bigram {“good point”}. The pre-processing removes stop words and extracts the mono-grams, bi-grams, and tri-grams.

In pre-processing I used $n=3$ which means using mono-grams (single words), bi-grams (two words), and tri-grams (three words) as the features extracted from the text. There is debate as to whether it is better to use mono-grams (set $n=1$) only or if Bag-of-Words classifiers should also use bi-grams and tri-grams (Pang & Lee, 2006). One trade-off is that the higher the n threshold the longer it takes to process text for the classification process. As the focus of Study 1 is on the creation of a classifier to increase the accuracy of prediction the higher and more computationally costly threshold value of 3 is selected. While increasing n when generating n -grams is known

to increase accuracy, 3 is considered an upper limit at the inclusion of four grams ($n=4$) does not increase accuracy (Thelwall, 2018).

After generating n -grams I then weighted the features using term frequency inverse document frequency (TF-IDF) which considers the frequency of a feature occurring within a text messages and divides the frequency within the message of the feature by the occurrence of the feature across all messages. TF-IDF weighs features by producing higher scores for unique features which has shown benefits to sentiment analysis (B. Liu, 2010).

The result of pre-processing is a document term matrix which is a matrix of features (n -grams) and associated weighted scores (TF-IDF). Given that I used student sourced labels the document term matrix also has an associated vector of the label provided by students. By using the document term matrix and associated labeled I train machine learning classifiers to detect the sentiment categories of positive negative neutral and mixed. While the decisions detailed for the use of n -grams, TF-IDF, and Noise words were in Study 1, I also examine all possible permutations of pre-processing for these decisions in Study 1.

3.4.6.2 LOGISTIC REGRESSION

For each valence category of positive, negative, and mixed using the Logistic Regression library in SciKit-Learn (Pedregosa et al., 2012) and training on student sourced labels we train SSSAC Logistic. This approach was aligned with recommendations by Shickel et al. (2016) that logistic regression works well for a four class sentiment analysis problem. This method used the liblinear solver, a library for large scale classification (Fan, Chang, Hsieh, Wang, & Lin, 2008), to handle the high dimensional data. Liblinear has shown promise in conjunction with logistic regression in sentiment analysis studies (Xia, Xu, Yu, Qi, & Cambria, 2016). Given that there was a class imbalance of examples in every experiment, the parameter `class_weight='balanced'` provided random oversampling. Random oversampling replicates the smaller class so that there is an equal number of records to compare and is a strategy used with machine learning research in education (Bosch et al., 2015).

For this study, this resulted in three classifiers that were combined into an ensemble to classify messages as positive, negative, or mixed. The classifiers ran in the following order: mixed, negative, positive. The first classifier that predicted the message belonged to the category was used to classify the message. If none of the classifiers predicted that a respective message belonged to one of three valence categories, then the message was classified as neutral. This process resulted in a classification of messages as Mixed, Negative, Positive, or Neutral. While this order was used to generate predictions used in Study 1, I also examine all possible permutations of processing order in Study 1.

3.5 DATA ANALYSIS

3.5.1 ANNOTATED DATA

In this thesis I take the position of CTE which defines emotion as collective intentionality where participants in the context are required to understand emotional expression. To pursue this goal, I collect labels from students on their own conversations which aligns with the common definition that SA seeks to model the intent of the author and the reaction it elicits in the intended audience.

I contrast student labels with labels generated by raters on Mechanical Turk (MTurk) which is a crowdsourcing method frequently used to generate labels for sentiment analysis research. While MTurk is designed to benefit from the wisdom of the crowd the annotations come from an external group. By contrasting student labels with MTurk labels we examine a benchmark of reliability of a crowdsourcing approach that is not comprised by students providing an external crowd which theoretically falls under a collective perspective, more closely aligned with collective intentionality, but missing the critical feature annotators coming from the context where the text annotated was written.

3.5.1.1 STUDENT SOURCED LABELS

To generate a set of labelled messages in the categories of positive, negative, neutral, and mixed I first asked students to provide samples of communications from each category. When considering how to make sentiment analysis for a specific domain

one approach is to get a small sample of data from the domain and use it to train classifiers (Yadollahi et al., 2017).

There is in fact many examples where emotional labeling is done through a crowdsourcing manner (Hutto & Gilbert, 2014; Morris & McDuff, 2009; Warriner et al., 2013). Some critics have pointed out that crowds may not produce reliable results (Hupont, Lebreton, Maki, Skodras, & Hirth, 2014). Given the ontological perspective that emotion is a collective intentionality (see section 3.2 for a full description) it would make sense that the best possible crowd to use when crowd sourcing emotional labels would be members of the social setting where the text messages were originally created. In the context of schools that would mean students would be the ideal crowd. As the literature review indicated that naïve coders at a Scottish University had a good level of agreement on rating texts (Gill et al., 2008). In addition, one advantage of using untrained coders is that a lack of better from their intuition (Waldinger et al., 2004). This indicates that crowd sourcing emotion labels for text messages is a reasonable approach to establishing the ground truth to evaluate SA technology. The primary consideration for research design is to consider replication as it is unclear how stable the approach would be to crowd source emotion labels with college students.

Students review their own group discussions and for each valence category they are were asked to provide estimates of the number of messages in their own discussion. If the estimate was above 0 they were asked to provide between one and three examples (see appendix 4 and 5 for complete instructions). As group discussions were comprised of multiple students this method generated multiple labels for the same message when more than one student used the same message as an example. From the examples I used the Expectation Maximization (EM) algorithm to establish ground truth.

3.5.1.2 MECHANICAL TURK LABELS

I used Mechanical Turk, a platform designed for many crowdsourcing activities including the generation of sentiment analysis labels, to establish crowdsourced valence labels. I paid raters on MTurk 0.07\$ U.S. Dollars to rate messages which aligns with similar rates for this task. Each message was rated by five people because the experimental design had students working in groups of five and the highest number of

ratings that occurred with student labels was five students. No exclusion criteria were used to exclude any ratings from MTurk, just as no students were excluded. The qualification of masters was employed ensuring that raters had previously been approved for 95% of the ratings they previously provided over a variety of tasks. To establish ground truth from the crowd sourcing method I used the Expectation Maximization algorithm (Dawid & Skene, 1979) to select the best possible label for messages rated on MTurk.

3.5.1.3 COLLECTIVE INTENTIONALITY, GROUND TRUTH, AND THE EXPECTATION MAXIMIZATION ALGORITHM

After students have provided labels the result is a set of messages where there are either one label or multiple labels. To select the “best” label in the case where multiple labels have been provided I use the Expectation Maximization (EM) algorithm (Dawid & Skene, 1979) which selects the best label considering a majority rules perspective to seed an iterative process. Once labels are selected based on majority rule the accuracy of each rater is computed for each category (i.e., positive, negative, neutral, and mixed). Then the prevalence of each category is computed. Using the prevalence and accuracy measures the algorithm iterates and revises labels. I use this selective process to model collective intentionality.

3.5.1.4 EXAMING THE EFFECTS OF EMOTIONAL SENTENCE STARTERS

In the second experiment, conducted in 2017, students are randomly assigned to either a control condition (2017C) which replicates the experiment conducted in 2016 or they are assigned to an emotional sentence starter (ESS) condition (2017SS) where they use scripting supports to explicitly state their reaction in terms of valence at least twice during group discussions. The intent of the intervention is to examine if ESS can operate as an emotion awareness tool to improve students’ ability to identify emotion expressions from their own group discussions. ESS asks students to explicitly state the valence of their reaction in text making the identification of valence in text trivial for the scripted statements. We examine the *unscripted* statements in this condition to examine the effect ESS has on students ability to identify valence in text.

3.5.2 RELIABILITY

To examine the reliability of annotations we first examine the examples students provided using Krippendorff's alpha to evaluate the agreement of annotation of the valence classes of positive, negative, neutral, and mixed because the statistic can be computed when there is unequal number of ratings per message annotated. Krippendorff's alpha can be weighted with a distance function between ratings, but previous work in sentiment analysis with categorical labels used an unweighted calculation (Chakravarthi et al., 2020) and I follow this approach. While Krippendorff's alpha can analyze unequal number of ratings it also works when there is a consistent number of ratings (Krippendorff, 1980). I apply the same overall agreement statistic to the researcher labels and the MTurk labels.

3.5.3 EVALUATING THE CLASSIFIER

3.5.3.1 CROSS-VALIDATION

For the machine learning classifiers nine of the folds were used to train the classifier, and the tenth fold was used to test the accuracy of the classifier. Each fold was used to test the classifier meaning that I calculated the accuracy of classifiers. Effectively cross-validation uses 90% of the data to train the classifier and 10% of the data to test the predictions of the classifier. By repeating this process 10 times, each fold comprised of 10% of the data is used to test the classifier trained on the other 90% effectively testing the classifier 10 times (see Figure 3.5.3).

When using 10-fold cross validation the recommended comparison for overall comparison is to add the predictions for each fold together, and then calculate the F-Score on the summation of the folds, as comparing mean scores has a bias to favor algorithms that produce more false positives (Forman & Scholz, 2010). For Study 1, F-Score is used to examine the overall accuracy, because F-Scores integrate False Positive (FP) and False Negative (FN) across all the predicted categories (i.e., positive, negative, neutral, and mixed) into one score. While F-scores provide a single comparable score

across measures for overall accuracy they do not provide details about how the performance of classifiers differs in terms of the categories predicted.

To understand the performance for each category it is suggested to have a separate calculated score that considers FP and FN (Stapor, 2017). I use recall as a measure that includes FN as recall is defined as a ration of true positives (TP) and false negatives (i.e., $\text{recall} = \text{TP}/(\text{TP}+\text{FN})$). I use the Polarity Bias Rate (PBR) (Iqbal, Karim, & Kamiran, 2015) as a measure that considers FN as PBR is defined as $(\text{FP}-\text{FN})/\text{Total}$. Effectively PBR determines is the appropriate amount is classified by quantifying if there is a bias towards predicting a specific label. Recall and PBR provide measures that consider FP and FN.

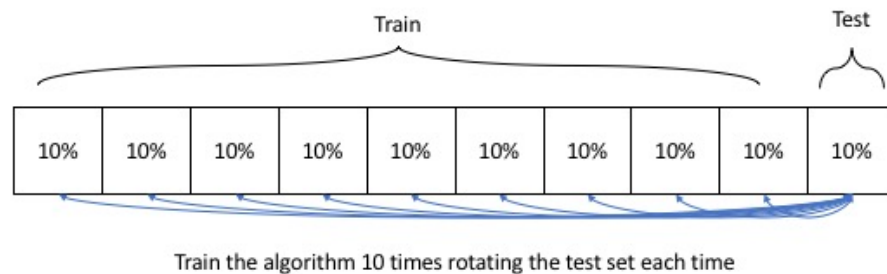


Figure 3.5.3 10-Fold Cross-validation

While the proposed cross-validation using student provided examples of positive, negative, neutral, and mixed messages can provide us with a sense of the extent to which a predictive algorithm agrees with student perception it does not confirm how that perception aligns with our theoretical model of emotion. From the CTE perspective peer perception is a critical element as consensus between members of a social group is what defines emotion. However, the components of emotion I discussed previously (appraisal, physiological response, action tendencies, and expression, and experience) are categorized by the CTE as three components that occur individually and two components that occur socially. While internal validation can consider alignment between student perception and a predictive algorithm of emotion in text, I can consider how measures that focus on other components of emotion align with this perception as a means for checking the validity of the measure with alternate measures.

3.5.3.2 TESTING CLASSIFIERS ON NOVEL DATA

While cross-validation provides a predicted level of accuracy it is important to also test the classifier on novel data (I.e., data not used to train the classifier). In this thesis we conduct a replication of the study in 2016 and use the 2016 data to train a classifier and use data from the replication as novel data to test the accuracy of the classifier.

As defined in section 3.4.6.1 Pre-Processing a number of decisions were made in terms of how to pre-process text prior to training a classifier. The initial configuration used n-grams including mono-grams, bi-grams, and tri-grams and removed noise words from the monograms. After generating n-grams the pre-processing applied the TF-IDF approach to weight text features. The algorithm we trained integrated three logistic regressions (mixed, negative, and positive) where the first classifier to make a true prediction was used. We also test this assumption by running a battery of tests which examines alternative orders for valence predictions using the best performing set of pre-processing features.

It was necessary to make decisions a priori to conducting the replication study as we needed predictions to use in the student interviews (see section 3.5.5). With both the training data collected in 2016 and the test data generated in the replication in 2017 we ran through all possible permutations of using one to three n-grams, removing noise words for each of the n-grams used, and weighting features using TF-IDF (yes/no). We report cross-validation scores from the training data and accuracy scores when using the classifier on novel data. We run these permutations using the three types of labels (student sourced, and MTurk labels) to see which ground truth has the best possible accuracy on novel data to evaluate which approach can produce the highest level of accuracy when making predictions on novel data.

3.5.3.3 STUDENT INTERVIEWS

While testing the classifier on Novel Data provides one approach to validating the predicted accuracy it does not provide any insight into potential disagreements between student ratings and predictions by the classifier. To explore disagreements, we conducted interviews with students to compare how they label messages with predictions from the algorithm. Specifically, we examine disagreements and ask students if the prediction of the algorithm changed their mind about the best valence label for text messages. The purpose of interviewing students is to remain close to the aim of the classifier which is to model student opinions.

For the student interviews we conducted one-hour semi-structured interviews and focus our analysis on three components of the interview. First, students were asked to rate messages from their own group chats in the categories of positive, negative, neutral, and mixed. Second, we compared the ratings students provided during the interview to the predictions from the classifier and asked students to identify if they agreed with the classifier (yes/no). If the student disagreed with the prediction of the classifier we asked students to look at the text features the classifier used to make a prediction and consider the text features and prediction to determine if the classifier changed their mind to agree with the prediction as the best possible label rather than their own ratings. Third, we asked an open-ended question to see if they found the predictions useful.

3.5.4 CORRELATION ANALYSIS

Triangulation is using different data collection activities including different instruments, times, or respondents to offset the weakness of in each single data collection by counterbalancing with another measure (McKenney & Reeves, 2014). When considering emotional measures it is generally advised to consider triangulation (Mauss & Robinson, 2009). As our review of measures related to the UME model of emotion indicated the Berkley Expressivity Questionnaire (BEQ) can measure dispositions for emotion expression as a bivariate measure, react can be used as a self-report mechanism interpretable in a bipolar manner. The Mixed Emotion Scale (MES) provides univariate mixed as a single dimension of mixed through students self-

reporting after an experience the extent to which they experienced mixed emotions. Similarly, the PANAS can measure the extent to which students experience positive affect and negative affect as a bivariate measure. By using these measure, I examine the extent to which a univariate mixed model of emotion expression (predicting the category of positive, negative, neutral, and mixed) correlates with measures using bipolar (positive, negative, and neutral), univariate mixed (mixed), and bivariate paradigms (positive, negative) across emotional state and trait.

To conduct correlation analysis between emotional measures, I administer the comparison measures (BEQ, React, PANAS, and MES) and then conduct confirmatory factor analysis where appropriate (BEQ, PANAS, and MES). When instruments fail confirmatory factory analysis I conduct exploratory factor analysis to select the best interpretation of items to test correlation between psychological measures of emotion and SA.

In this PhD thesis I focus on the correlates with measures designed to detect individual emotional states and traits for a measure that is based on emotional communication. From our theoretical perspective using the CPM diagram (see Figure 2.1.1) not all communication is considered a Valid Self-Report of Emotion. In fact, I detailed the consensual model of emotion expression to articulate why and how communication can be disassociated from emotion (see section 2.1.1.2). I also categorized two additional forms of communication (Regulated Communication, Disconnected Communication) as two types of communication that is disassociated from emotion (see Figure 2.1.1). In part this is why I included a measure of traits for emotion expression to help consider the limitations of correlation between emotion self-reported before during and after the activity with expressions in the activity. While there is reason to believe that not all communication would exhibit correlation with additional measures of emotion the focus of correlation analysis focuses on the extent to which emotional communication in online group work demonstrates a correlations with psychological measures of emotion.

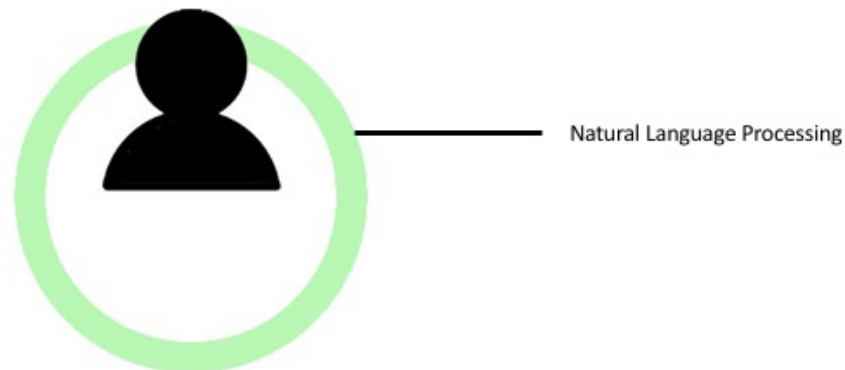
3.6 ETHICS

An ethical review board approved the study protocol and all study related documents including a consent form, and the post activity. The consent form was given to participants at the start of the group work activity. This work had ethical approval by the Open University with HREC/2016/2388/Hillaire/1. A detailed description of the ethics application is available in the Appendix 6 and Appendix 8.

3.7 CONCLUSION

This chapter outlined the methodology of this PhD thesis, identifying methods and ethical considerations. The next three chapters detail studies in depth, Chapter 4 describes Study 1, Chapters 5 describes Study 2, and Chapter 6 describes Study 3.

CHAPTER 4 STUDY 1 – INTRODUCING STUDENT SOURCED SENTIMENT ANALYSIS



4.1 INTRODUCTION

In the first empirical Study 1, I focus on Sentiment Analysis (SA) which is commonly defined as the detection of how the opinion of the author of the text elicits a reaction from the intended reader of the text (Balahur & Steinberger, 2009). More specifically I introduce a crowdsourcing method where students examine their own online chat messages and provide examples to train a classifier. Students first engaged in an online group discussion in small groups and then reviewed their group discussion and selected examples of messages for the valence categories of positive, negative, neutral, and mixed. The examples were comprised of messages they wrote themselves and messages written by peers which aligns with the definition of SA as opinion of the author and reaction from the intended reader. Most crowdsourcing approaches start with the assumption that individual ratings are noise until multiple ratings agree to establish a ground truth (I.e., correct label). This student sourced sentiment analysis classifier (SSSAC) approach takes the opposite perspective that the crowd is not an anonymous group disconnected from the context, but rather the people whose opinions I am aiming to model.

SA is a common and established approach towards detecting emotion in text expression (Pang & Lee, 2006). For example, a review of affective computing (AC), which is a branch of computer science that aims to recognize and respond to emotional states, Calvo and D'Mello (2010) describe SA as usually representing words in multi-dimensional space (MDS) to categorize text into dimensions of emotion (e.g., the dimensions of valence). When applying AC to the context of education, it is referred to as affective learning (AL), which investigates how emotions affect learning based on the perspective that some affective states facilitate different kinds of thinking than others, and different kinds of thinking have long been important to research on learning (Picard et al., 2004). (For a detailed review of SA in AC and AL, see section 2.1).

As already indicated in section 2.1, there is already a lot of promising evidence in AC that SA can help to understand the role of emotion in text expression (B. Liu, 2010; Pang & Lee, 2006; Thelwall, 2013). There is further promising evidence in AL that SA can help explore how emotion expression in text relates to learning (Lang et al., 2017; Rienties & Rivers, 2014). When reviewing 15 studies that used SA in the context of learning (See Table 2.1), there were many promising results. For example, some research found correlations between SA in online courses and student retention (Chaplot et al., 2015; S. Crossley, Paquette, et al., 2016; Wen et al., 2014). SA can also be used in conjunction with self-report to gain insights into the student experience while learning (Calvo & Kim, 2010; Rajput et al., 2016; Santos et al., 2013). Furthermore, SA researchers are starting to explore how to highlight for students the emotion expressed in online chat (Ortigosa et al., 2014).

SA research shows promise regarding investigations into the complex role of emotion in learning. Given the potential for SA in educational research, it is essential to consider the validity and reliability of SA. When reviewing 15 studies of SA in the context of AL, I found that only three studies (Calvo & Kim, 2010; Rajput et al., 2016; Santos et al., 2013) evaluated the accuracy of their proposed measure based on the perspective of authors (i.e., students). None of the studies used the perspective of the intended audience (see Table 2.1).

Therefore, I propose that by having students participate in a group chat and reviewing all the messages from the chat, they can provide a direct evaluation of the

emotion in the text as both authors and intended readers of the text. When reviewing their own messages, as the author they may have insights into the opinion they expressed. When students are reviewing messages written by peers, they are the intended audience, and may have insights into the reaction elicited by the text in the social context of the classroom. While there are potential benefits to asking students about their perspective on emotional communication, there is also the consideration that people may be limited in their ability to even be aware of their own emotions. In part this may be because the affective state is geared toward the present, making people potentially unreliable historians of their past affective states (Pham, 2004). While there are potential benefits in asking people to identify emotion in their own communication through reflection on their own writing, there are limitations. Therefore, the purpose of Study 1 focuses on the extent to which measures align with students' ability to recall what emotion was expressed in their group discussions as a reflection exercise after the lab activity.

As I use a crowd sourcing method it is suggested to examine the agreement of the crowd to understand the reliability of their ratings. Specifically reporting Fleiss' kappa and Krippendorff's alpha is suggested best practice which I explore in research question 1 (RQ1): To what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters? RQ1A: To what extent do students agree in terms of inter-rater agreement when providing examples? RQ1B: To what extent do Mechanical Turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples?

In order to incorporate the student perspectives into a SA measure, Study 1 uses the labelled messages provided by 767 students to train the machine learning classifier SSSAC Logistic. When asking students to self-report emotion there is a tension between using familiar terms and validated measures as indicated by a ten-year review in the journal *Emotion* (Weidman et al., 2016). This is supported by the perspective that one of the common breakdowns in self-report of emotion is using terms used in an instrument must be familiar to the people providing the self-report (Duckworth & Yeager, 2015). In fact, a recent review of self-report approaches in the journal *Emotion* indicated that 246 out of 356 (69%) measurements used impromptu scales, which may allow flexibility for

researchers to use terms familiar with the research subjects, but can make it hard to compare findings across studies (Weidman et al., 2016). The recommendation for emotion researchers is to avoid using impromptu scales and rather build on existing research and theory by creating self-report approaches that models the entire universe of the construct under investigation (Weidman et al., 2016).

This leads us to the field of psychology to identify which emotion labels might be both validated and familiar for students, which reflects the totality of the construct under investigations, when reflecting on emotion expressed in their own text. One of the earliest ways psychology researchers categorized emotion is to consider if the emotional experience was positive or negative, in part because this is considered a very intuitive way to categorize emotions (Blumenthal, 1975). When using SA with the intended goal of supporting self-reflection, the proposed ideal from the field of psychology is the detection of four categories: positive, negative, neutral, and mixed (which is both positive and negative) (Shickel et al., 2016). These four categories make up valence, which can be conceived as either a dimension from negative to positive (Russell & Carroll, 1999), or as a plane of positivity and negativity (Cacioppo et al., 1999). For a detailed discussion of valence, see section 2.2.

The plane of positivity and negativity is constructed with positive as the x-axis, negative as the y-axis, neutrality as the origin, and mixed valence in the plane of positivity and negativity referred to as the evaluative space model (ESM) (Cacioppo et al., 1999, 2004). When interpreting emotions that are both positive and negative I use a univariate mixed paradigm, as described in detail in section 2.3, which directly asks students to categorize text messages as positive, negative, neutral, or mixed. By selecting a valence model ideal for self-reflection, it has theoretical advantages for integrating the perspective of the authors and intended audience of the text to determine the accuracy of SA. The students themselves determine which of the four valence categories best describes text messages, which may require considering messages that are both positive and negative, and selecting which valence category is the most appropriate. Because the individual students are making these judgment calls, this process classifies as a univariate mixed paradigm. One rater determines if something that is both positive and negative should be considered overall positive, overall negative,

mixed, or if the two positive and negative aspects neutralize (Kreibig & Gross, 2017). For more details on the univariate mixed paradigm see details in section 2.3. For Study 1 the implications is that I ask students to classify messages in four categories (i.e., Positive, Negative, Neutral, Mixed) that according to the ESM represent the totality of the construct of valence (Cacioppo et al., 1999, 2004). This raised research question 2 (RQ2): To what extent can crowd sourced, and in particular student sourced, examples train a machine learning classifier to predict the valence categories of positive, negative, neutral, and mixed? RQ2A: To what extent can student labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2B: To what extent can Mechanical Turk labels train a logistic classifier which predicts the valence categories of positive, negative, neutral, and mixed? RQ2C: How do logistics classifiers trained using student labels and Mechanical Turk labels compare to general benchmarks when predicting the valence categories of positive, negative, neutral, and mixed? RQ2D: To what extent do students find predictions from a student sourced classifier useful?

In summary, Study 1 follows the guidance to build on existing emotional measurement work (Weidman et al., 2016) by aiming to improve SA, which can predict the dimension of valence (Calvo & D'Mello, 2010; Pang & Lee, 2006). From a theoretical perspective of ESM, the entire universe of the construct of valence includes the categories of positive, negative, neutral, and mixed (Cacioppo et al., 1999, 2004). These four valence categories are considered ideal to support self-reflection of authors examining their own text (Shickel et al., 2016). When asking students to review their own text messages and code the text as positive, negative, neutral, or mixed, the aim is creating a measure that follows the univariate mixed paradigm (Kreibig & Gross, 2017). This study takes the initial step towards validating the measure SSSAC Logistic by examining the inter-rater reliability of student sourced labels (RQ1) as well as the accuracy of the measure for the valence categories of positive, negative, neutral, and mixed (RQ2).

4.2 METHODS

In 2016, the format of the Pilot Experiment focused on the generation of a student sourced univariate mixed SA classifier SSSAC Logistic. Building on previous research conducted in an authentic computer-based environment at one business school in the Netherlands (Knight et al., 2017; Mittelmeier et al., 2018) which found that this online collaborative task was appropriately designed for encouraging active engagement across 428 students in a Randomized Control Trial.

4.2.1 SETTING

The study took place at a University in the Netherlands in a freshman statistics course. The University recruited international business students as part of the teaching philosophy at the University was that students could learn from a diverse group of peers. In this context students typically had a problem-based learning (PBL) curriculum, meaning that they were used to working in groups to solve a specific problem, see Tempelaar et al. (2015; 2017) for a detailed description of the educational context in which Study 1 collected the data. As described in greater detail in Mittelmeier et al. (2018), the World-Bank assignment asked students to discuss data from a set of countries and work on a problem of making a funding decision in a group. The group nature of working on a problem together made this assignment ideal for this context. The primary distinction between the lab-based assignment and common experience of students was the use of an online platform, Udio, to facilitate the group work via an online chat interface.

4.2.2 2016 PROCEDURE AND PARTICIPANTS

In the 2016 procedure, students were assigned randomly to groups of 5 ($M=4.73$ $SD=0.84$) in a laboratory setting, whereby each student had a desktop computer, and all written communication was online as part of a regularly occurring lab session for their course. The discussions took place in Udio, which is a platform designed to support reading comprehension through providing short high interest content with integrated reading comprehension supports, such as discussing the reading (Hillaire et al., 2018). In Study 1, Udio delivered the group assignment and supported group discussion. The

average number of comments per group was 80 (M=80.69 SD=27.07). Previous research in this context has highlighted that the lab environment was appropriate for online experiments in collaborative learning, and a recent study has found that these students enjoyed the authentic World Bank task, and overall students enjoyed working together in groups (Mittelmeier et al., 2018).

Students participated in a group exercise using a Udio platform during a lab where they first introduced themselves and offered fun facts about themselves as an icebreaker. Then the students discussed an authentic collaborative task using World Bank statistics about educational achievement. Finally, in the lab they provided a single group response to the question “which country that your group was discussing deserves additional funds to achieve the goal of having everyone enrolled in higher education?” Previous research (Mittelmeier et al. 2018)

After leaving the lab exercise students were asked to complete a post activity within a week of the lab. In the post activity students logged back into the Udio platform used during the lab (see Figure 4.2.2.) and reviewed their group discussion comments to provide examples of positive, negative, neutral, and mixed chat messages from their own group chat. These messages were then used to train the SA classifier SSSAC Logistic capable of identifying if a message is positive, negative, neutral, or mixed. SSSAC Logistic is considered a SSSAC because the data set used to train the classifier was generated by students using crowd sourcing methods. SSSAC Logistic is also considered a univariate mixed SA technology because students directly evaluated if the messages were positive, negative, neutral, or mixed.

Process			Output
Ice Breaker (5 min)	Discuss World Bank Data (50 min)	Provide an Answer (5 min)	Post Activity Label Messages (60 min)
During Lab			After Lab

Figure 4.2.2 Pilot Experiment Conducted in 2016

The 2016 Data Collection was comprised of 767 first-year business students took part in Study 1 during Week 6 of instruction. The mean age was just under 19 ($M=18.95$; $SD=1.28$). There were 304 females and 463 males. The population was international, including 191 domestic, 529 European Students, and 47 non-European students.

4.2.3 2017 PROCEDURE AND PARTICIPANTS

The 2017 procedure took place during Week 6 of instruction in the same setting and course as described in the 2016 participants, 447 students were assigned randomly to groups of 4 ($M=3.49$; $SD=0.89$). The mean age was just over 19 ($M=19.01$; $SD=1.14$) and comprised of 178 females, 262 males, and 7 unknown gender. Using the same activity described in 2016 with a few changes.

The first change was that the initial icebreaker activity to familiarize students with Udio and the respective group they were working in during the lab study was replaced with a brief statistics sampling activity. This was specifically included as requested by the teacher of the respective mathematics and statistics module, as students in 2016 at times could not always make a link between the statistics course and the World Bank lab activity.

The second change in comparison to 2016 was that students were instructed to self-report their emotion and view the emotional reactions of their peers using the React tool in Udio. This change was designed to get a specific emotional measure of incidental emotions that students felt at the beginning of the group exercise. Figure 4.2.3a visually illustrates how students could self-report their emotional reactions using a multi-select of twelve emotion words. React provided the following list of words as options: engaging, interesting, challenging, curious, calming, good, dull, boring, sad, confusing, frustrating, and annoying. In Figure 4.2.3a the words engaging, interesting, challenging, good, boring, and sad have been selected as an example of what the student would see after making a selection. While the responses of React are not analyzed in this study the change is reported to describe completely the change in procedure.

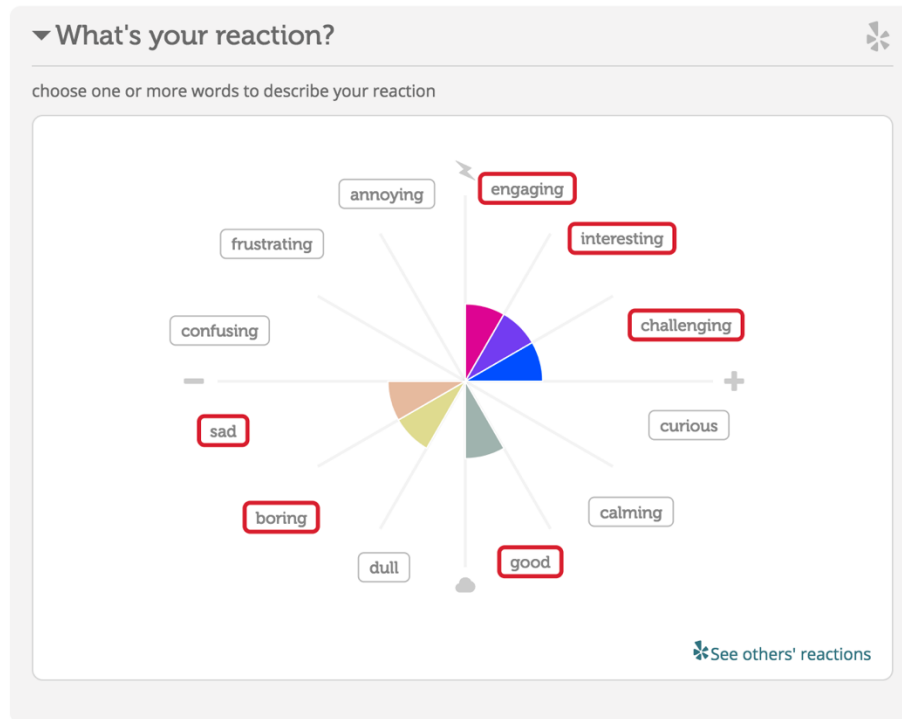


Figure 4.2.3a - Interface of 'React' to self-report emotional response

The third change to the process was to remove the group response component and move the answer to the case study into the output of the experiment by having students provide individual answers during an exit survey. A detailed breakdown of the Procedure is provided in Figure 4.2.3b.

Process			Output			
Sampling Activity (10Min)	<u>Self-Report</u> React (5 min)	Discuss World Bank Data (30min)	Provide an Answer (10min)	<u>Exit Survey</u> MES PANAS (10min)	Post Activity Label Messages (60 min)	<u>Interview</u> Accuracy Aggregate Evaluation
During Lab					After Lab	

Figure 4.2.3b Input, Process, Output for Main Study conducted in 2017

4.2.4 INSTRUMENTS

4.2.4.1 STUDENT SOURCED EXAMPLES

In the post-lab exercise, I asked students to review messages from their own group discussions and provide examples. The primary incentive for students to participate in this part of the study was that they were offered the option to either do their usual lab activity or participate in the study instead of the doing the lab work. This resulted in a high level of participation of students in the study. For each valence category of positive, negative, neutral and mixed, students were asked to identify if there were more than zero messages. If they saw more than zero messages for a given category they were then instructed to provide between one and three examples. As I were using a crowd sourcing method, where students selected the examples of emotional expression in their text messages, there were inevitably some examples where multiple students provided the same message as an example, where the label they provided for the message might (not) agree with the label provided by their peers. Therefore, the examples provided by the students were pre-processed using the expectation maximization (EM) algorithm (Raykar et al., 2010) which uses simple majority, frequency, and accuracy of raters to determine the most accurate label in situations where multiple raters provide ratings for the same message. The result of this was a set of unique messages with a correct label that was either positive, negative, neutral, or mixed (for a full description of the EM algorithm see section 4.2.5.1).

In order to address RQ1 and RQ2 and compare and contrast our SSSAC Logistic SA approach in terms of accuracy, I included 10 common heuristic and lexical SA instruments, as well as benchmark against a classifier trained on mechanical turk labels. I will briefly describe these approaches in sections 4.2.4.2-5.

4.2.4.2 HEURISTIC SENTIMENT ANALYSIS INSTRUMENTS

Using the student sourced examples (see section 4.2.4.1) I used heuristics benchmarks, which are simple rule-based approaches to predict the correct label of the messages category. When using heuristic benchmarks to gauge the accuracy of student

sourced SA our review indicated that one AL study (Calvo & Kim, 2010) used two heuristics focused on the student experience which were Random Baseline (RB), which is randomly guessing the correct label (e.g., when predicting positive, negative, neutral, and mixed each category would be guessed ~25% of the time), and Majority Class Baseline (MCB), which is always guessing the most commonly occurring label (e.g., if positive is the most common then this heuristic would always predict positive).

4.2.4.3 LEXICAL SENTIMENT ANALYSIS INSTRUMENTS

For Study 1 I used bivariate SA technologies and the co-occurrence method (which is bivariate mixed paradigm) to categorize text as positive, negative, neutral, and mixed. Séance, Vader, and LIWC are technologies that are included in lexical SA instruments because in our review they have been used in AL studies. SentiStrength was also included in Study 1 because the technology was created referencing the ESM of valence, which is the same model referenced when creating SSSAC Logistic.

Séance provides a single interface to use a variety of freely available SA technologies. I also used the co-occurrence when interpreting the bivariate predictions from Séance. For Study 1 I used Séance to generate bivariate predictions using GALC, GI, EmoLex, Hu-Liu, and Lasswell. GALC, the Geneva Affect Label Coder, attempts to categorize 36 affective label experiences and is capable of providing a positive and negative score that reflects the detection of those affective experiences regarding valence (Scherer, 2005). GI, the General Inquirer, includes the Harvard IV-4 dictionary lists manually constructed containing over 11,000 words, and is the oldest list in widespread use (S. A. Crossley et al., 2017). EmoLex is a crowd sourced SA technology which predicts discrete emotions as well as bivariate valence developed for general purposes (Mohammad & Turney, 2010, 2013). EmoLex also used part of the General Inquirer for its creation (Mohammad & Turney, 2013). Hu-Liu uses a list of over 2,000 negative words and over 4,000 positive words mined from product reviews (S. A. Crossley et al., 2017; Hu & Liu, 2004; B. Liu & Street, 2005). Lasswell is a lexicon of 63 word lists organized into nine categories (power, rectitude, respect, affection, wealth, wellbeing, enlightenment and skill) (S. A. Crossley et al., 2017). Lasswell is also included in the General Inquirer (Lawrence & Rodriguez, 2012). While Séance is

capable of making predictions based on VADER, for Study 1 I used a more recent version of VADER than the version included in Séance.

VADER is an example of a SA technology that used a heuristic approach (Hutto & Gilbert, 2014), meaning that heuristic rules capable of predicting the valence of text are used to create the technology. VADER claims to have higher than human levels of accuracy (Hutto & Gilbert, 2014). The result of a VADER prediction is three probabilities of valence. A probability for Neutral, Positive, and Negative. VADER results were interpreted as a bivariate prediction about an emotion present in text, meaning by using the probability of positivity and the probability of negativity. When either probability was greater than zero it was interpreted to indicate that positivity and/or negativity was present.

LIWC, the Linguistic Inquiry and Word Count is a dictionary approach to SA. The technology is based on dictionaries that are composed of almost 6,400 words, word stems, and selected emoticons. LIWC is a common benchmark technology which was even used to benchmark Séance (S. A. Crossley et al., 2017). LIWC provides a bivariate prediction about the emotion present in text meaning that it produces a score for positivity and a score for negativity.

SentiStrength, is a tool from AC which was included in Study 1 because this SA technology was built referencing the ESM perspective on valence (Thelwall, 2013; Thelwall et al., 2010). SentiStrength provides a bivariate prediction about the emotion present in text, meaning that it produces a score for positivity and a score for negativity. The positivity score is an integer from one to five with one indicating no positive emotion and 5 indicating very strong positive emotion (“Twitter Emotion Coding Instructions,” 2013). Scores between two and four indicate some positive emotion that is not very strong (“Twitter Emotion Coding Instructions,” 2013). The same instruction are used for the scores of negativity (“Twitter Emotion Coding Instructions,” 2013).

All of the lexical bivariate prediction scores were interpreted using the co-occurrence method which considered the presence of positive and negative to be mixed valence (Larsen et al., 2017) resulting in a prediction that the messages were positive, negative, neutral, or mixed (both positive and negative). The co-occurrence method is

classified as part of the mixed bivariate paradigm (Kreibig & Gross, 2017) (To see a full description of the theory of valence measurement see section 2.2.3).

4.2.4.4 MACHINE LEARNING SENTIMENT ANALYSIS INSTRUMENT

When reviewing AC studies of SA (see section 2.1.1) Shickel et al. (2016) recommended the detection of positive, negative, neutral, and mixed for valence detection and also found that the Bag-Of-Words classifier performed best in comparison with six benchmarking technologies. I use logistic regression to predict the valence classes of positive, negative, and mixed. With three classifiers trained I use the valence label from the first classifier processing them in order of positive, negative, and mixed. Effectively if the positive classifier produces a positive prediction then the negative and mixed classifier results are not used. If positive produces a false prediction then the negative classifier results are examined. Finally, if the mixed classifier produces a false prediction then I interpret the message to be neutral.

4.2.5 ANALYSIS

The approach for this analysis was to crowd source valence labels from the students evaluating their own group discussion comments as a reflection activity. The crowd sourced labels were first selected by the EM algorithm. The labelled messages were then pre-processed to extract features from the messages to train the machine learning classifiers. After training the classifiers 10-fold cross validation was used to test the expected accuracy of the machine learning classifiers, in line with recommendations by (Browne, 2000; Forman & Scholz, 2010; Little et al., 2017). Then novel data generated by a different group of students in 2017 is used to test the accuracy of the classifier.

4.2.5.1 CROWD SOURCED LABELS

I generated a set of chat messages with student labels and refer to these examples as student sourced examples (see 4.2.4.1).

4.2.5.2 PRE-PROCESSING TEXT

When determining features for a text classifier, it is common to select a threshold value n , and extract text features that are below n for n -gram parsing, where n -grams represent co-location of words that are n in length. For example, the phrase “This is a good point” has five monograms {“this”, “is”, “a”, “good”, “point”}; four bi-grams {“this is”, “is a”, “a good”, “good point”}; and three trigrams {“This is a”, “is a good”, “a good point”}. When parsing text frequently stop-words (commonly used words) are removed prior to computing features (e.g., n -grams) from text (Lonchamp, 2012; Wiebe et al., 2005). In the example “This is a good point” when removing stop-words the phrase becomes “good point” as the first three words are commonly occurring words. The phrase that remains after stop-word removal, “good point”, has two monograms {“good”, “point”}; and one bigram {“good point”}. The pre-processing removes stop words and extracts the mono-grams, bi-grams, and tri-grams.

In pre-processing n -grams I used $n=3$ which means I evaluation mono-grams (single words), bi-grams (two words), and tri-grams (three words) as the features extracted from the text. There is debate as to whether it is better to use mono-grams (set $n=1$) only or if Bag-of-Words classifiers should also use bi-grams and tri-grams (Pang & Lee, 2006). One trade-off is that the higher the n threshold the longer it takes to process text for the classification process. As the focus of Study 1 is on the creation of a classifier to increase the accuracy of prediction the higher and more computationally costly threshold value of 3 is selected. We use this selection to make predictions used in this Study and rigorously evaluate this decision by comparing accuracy using between 1 and 3 n -grams in section 4.3.3

4.2.5.3 PROCESSING TEXT

In our review the Bag-of-Words algorithm was suggested to perform best when predicting the four valence categories of positive, negative neutral, and mixed (Shickel et al., 2016). For this evaluation I will use all of these two baselines approaches, SVM and NB, as machine learning benchmarks to compare with the proposed SSSAC Logistic classifier. As all three classifiers can use the same data to train and test the accuracy, the text is pre-processed once and then the same resulting data is used for all

three classifiers. Each classifier was trained on the pre-processed text of mono-grams (single words), bi-grams (two words), and tri-grams (three words).

SSSAC Logistic is introduced as an implementation built using the library in SciKit-Learn (Pedregosa et al., 2012) for the logistic regression as this implementation uses the liblinear solver, a library for large scale classification (Fan et al., 2008), to handle the high dimensional data. As I anticipated class imbalance, the parameter `class_weight='balanced'` provides a random over sampling strategy to account for training bias, which might favor associations of features with a majority class over a minority class due to the frequency and infrequency of examples. Effectively the balanced feature randomly duplicates examples from minority classes until there is an equal number of examples in all classes prior to training the classifier. For Study 1 I created three classifiers that were combined into an ensemble to classify messages as positive, negative, or mixed. The classifiers ran in the following order: mixed, negative, positive. The first classifier that predicted the message belonged to the category was used to classify the message. If none of the classifiers predicted that a respective message belonged to one of three valence categories, then the message was classified as neutral. This process resulted in a classification of messages as Mixed, Negative, Positive, or Neutral. We use this order in Study 1 and rigorously evaluate the decision in section 4.3.3

4.2.5.4 TEN-FOLD CROSS-VALIDATION

When splitting the examples into ten folds (i.e., one tenth of the examples were assigned to a fold), I used a stratified sample approach meaning that I had roughly the same proportion of positive, negative, neutral, and mixed examples in each fold. For the machine learning classifiers nine of the folds were used to train the classifier, and the tenth fold was used to test the accuracy of the classifier. Each fold was used to test the classifier meaning that I calculated the accuracy of classifiers. Effectively cross-validation uses 90% of the data to train the classifier and 10% of the data to test the predictions of the classifier. By doing this 10 times each fold comprised of 10% of the data is used to test the classifier trained on the other 90% effectively testing the classifier 10 times.

When using 10-fold cross validation the recommended comparison for overall comparison is to add the predictions for each fold together, and then calculate the F-Score on the summation of the folds, as comparing mean scores has a bias to favor algorithms that produce more false positives (Forman & Scholz, 2010). For Study 1, F-Score is used to examine the overall accuracy, because F-Scores integrate False Positive (FP) and False Negative (FN) across all the predicted categories (i.e., positive, negative, neutral, and mixed) into one score. While F-scores provide a single comparable score across measures for overall accuracy they do not provide details about how the performance of classifiers differs in terms of the categories predicted.

To understand the performance for each category it is suggested to have a separate calculated score that considers FP and FN independently (Stapor, 2017). Recall, which is defined as a ration of true positives (TP) and false negatives (i.e., $\text{recall} = \text{TP}/(\text{TP}+\text{FN})$). PBR $(\text{FP}-\text{FN})/\text{Total}$ provide measures that consider FP and FN as two different measures.

4.2.5.5 COMPARATIVE ANALYSIS METHOD

To answer RQ1, I compute Krippendorff's alpha to evaluate student agreement on example labels and compare that with agreement of Mechanical Turk generated labels. To answer RQ2, I examine F-scores for machine learning classifiers trained on student labels as compared with a classifier trained on Mechanical Turk labels as well as general benchmarks.

RQ1 & RQ2 have a focus on accuracy of SA. To contextualize the results, I selected the conversation with largest number of student sourced labels to inspect in detail. While Chapter 6 has an emphasis on the validity of the measure, this conversation inspection is meant to provide an early indicator of face validity in the thesis. The intent of the example conversation is simply to provide context to help interpret the results of RQ1 and RQ2 by contrasting how SSSAC Logistic classifies text with SentiStrength. I used SSSAC Logistic and SentiStrength on the same conversation because both technologies are built using the ESM (see section 2.3 for further details) and both classifiers are human interpretable. This example provides an early indicator into how SSSAC Logistic works.

4.3 RESULTS

4.3.1 DATA COLLECTION AND GROUND TRUTH ESTABLISHED USING THE EXPECTATION MAXIMIZATION ALGORITHM FOR THREE DATASETS

4.3.1.1 2016 DATA COLLECTION (TRAINING DATA)

767 students were asked to complete a post-activity (see Appendix 4 for the full activity) after participating in small group online chats in a lab session (see 4.2.2 for a description of the lab). During the post activity students were asked to follow the instructions:

For the next set of questions, we ask you to look at all the messages in the conversation log in Udio. When looking at all the messages, try to identify which message contains a **positive** or **negative** reaction to the data from the World Bank. Not all messages will easily fall into positive or negative reactions, however. Some messages may lack emotion and can be considered **neutral**. Some messages may contain both positive and negative content and should be considered **mixed**. Some messages may be difficult to determine if they are positive or negative, and these messages can be considered **ambiguous**. Here are some examples of sentences that are positive, negative, neutral, mixed, and ambiguous:

- **Positive:** “I really like that the Netherlands contributes 5.5% of their GDP for education”
- **Negative:** “The Netherlands only contributes 5.5% of their GDP for education which is not enough”
- **Neutral:** “The Netherlands contributes 5.5% of their GDP for education”
- **Mixed:** “It is good that the Netherlands contributes 5.5% of their GDP for education, but unfortunately they do not spend the money wisely”
- **Ambiguous:** “The Netherlands contributes 5.5% of their GDP for education, that seems pickles to me.”

In total 16,590 messages were posted by 767 students, leading to a large dataset. In total 521 out of 767 students (68%) provided at least one example message on the post-activity. These 521 students provided a total of 2530 examples comprised of 1979 unique messages, which is 15.25% of the total text dataset. The examples they provided came from their own group discussions, so if two students in a respective group were examining the same discussion thread then they were independently selecting examples

for each valence category. There were 446 messages where more than one person selected the same message as an example, of which for 258 the same valence category was selected, while for 188 examples there was disagreement. Table 4.3.1.1a reports the number of students that provided the same message as an example.

Table 4.3.1.1a Number of Students Per Example

Number of Students	Messages (Unique)	Total Examples Provided (Not unique)
One	1533	1533
Two	359	718
Three	216	216
Four	12	48
Five	3	15
Total	1979	2530

The EM Algorithm used the simple majority to initialize the accuracy of the raters and the prevalence of the valence categories. With accuracy and prevalence values the labels were adjusted. After adjusting the labels, the accuracy and prevalence were recalculated. This iterative process converged with a single label selected by the algorithm for each message. The EM algorithm generated what Study 1 considered the best label to describe the student experience for each example message. For 388 examples the algorithm selected the label of ambiguous indicating that the algorithm indicated the most accurate label was that the message was ambiguous and not clearly in one of the categories of positive, negative, neutral, or mixed. When subtracting 388 from 1979 the remainder is 1778 unique examples

Table 4.3.1.1b Student Examples and Associated Labels

	Positive	Negative	Neutral	Mixed	Total
Unique	587	444	524	223	1778
Examples	33%	25%	29%	13%	

Table 4.3.1.1b outlines for the unique examples categorized by what EM determined to be the “correct” labels. This set of unique examples with valence labels are the standard by which we train a classifier Study 1. The majority class is Positive which has 587 examples (33%), followed by Neutral with 524 examples (29%). Negative is the third class with 444 examples (25%), while Mixed is the minority class with 223 examples (13%). Based upon the 1,979 unique messages labeled by students, Table 4.3.1.1c shows examples of messages provided by one and only one student as well as examples where more than one student provided the same message as an example.

Table 4.3.1.1c EM Selection of Student Labels

Label	Type	Examples	Comment
Positive	NC	1 POSITIVE	“yeah that would be great [FIRSTNAME] :)”;
Positive	NC	1 POSITIVE	“haha.”
Positive	EM	1 POSITIVE 1 NEUTRAL	“In the netherlands, you see that gender doesn't really have any influence on education. Woman and man are both +- equal”
Negative	NC	1 NEGATIVE	“Okay wow Guatemala is the worst”;
Negative	NC	1 NEGATIVE	“[Expletive].”
Negative	EM	3 NEGATIVES 1 AMBIGUOUS	“It's just surprising that between sexes males are still predominant on the primary enrollment and completion rate”
Neutral	NC	1 NEUTRAL	“In Italy the GDP is 4.1”;
Neutral	NC	1 NEUTRAL	“I would say Yemen”;
Neutral	EM	1 NEGATIVE 1 NEUTRAL	“the goverments expenditure in mali is only 4.3 %”
Mixed	NC	1 MIXED	“I'm confused on what's going on haha”;
Mixed	NC	1 MIXED	“for example they have been really bad in science but really improved over the years.”
Mixed	EM	1 MIXED 1 NEUTRAL	“Germany has more dropouts than Netherlands, and less than Greece”

The positive examples show students complimenting and agreeing with each other. The negative examples are illustrative of describing emotional reactions through the use of an expletive, an overall judgement about a country, and a judgement about educational achievement by gender. The examples of neutral communications appear to be more data-oriented. There are two EM examples, in Table 4.3.1.1c, that place a value judgement on a gender related comment, one which indicates in the Netherlands there is equality in achievement for males and females, while the other comment indicates that in some unstated location men are more likely than women to enroll in higher education. The mixed expressions demonstrate communications that include multiple viewpoints. “bad in science but really improved over the years” demonstrates two concepts (i.e., “bad in science” and “improved over the years”).

In the mixed examples, the first thing to note is that confusion is frequently mentioned. One example of a comment about confusion in Table 4.3.1.1c appears to represent students turning a potentially negative experience of confusion into a complex emotion by communicating “I’m confused” in conjunction with the positive addition of laughter (“haha”). The other examples appear to speak about two contrasting points that are positive and negative. It may be that reflexive mixed emotions are about complex emotions of learning while mixed emotions directed at the course material represent weighing pros and cons about the course material.

4.3.1.2 2016M MECHANICAL TURK DATA COLLECTION (TRAINING DATA)

We used the EM algorithm to select ground truth based on mechanical turk Labels resulting in the following composition of ground truth. We provided the 1778 examples determined to be in the categories of positive, negative, neutral, and mixed by students to Mechanical turk and asked raters to pick one of the four categories using the same instructions as students (removing the ambiguous option):

For the next set of questions, we ask you to look at all the messages in the conversation log in Udio. When looking at all the messages, try to identify which message contains a **positive** or **negative** reaction to the data from the World Bank. Not all messages will easily fall into positive or negative reactions, however. Some messages may lack emotion and can be considered **neutral**. Some messages may contain both positive and negative content and should be considered **mixed**. Here are some examples of sentences that are positive, negative, neutral, and mixed, and ambiguous:

- **Positive:** “I really like that the Netherlands contributes 5.5% of their GDP for education”
- **Negative:** “The Netherlands only contributes 5.5% of their GDP for education which is not enough”
- **Neutral:** “The Netherlands contributes 5.5% of their GDP for education”
- **Mixed:** “It is good that the Netherlands contributes 5.5% of their GDP for education, but unfortunately they do not spend the money wisely”

Table 4.3.1.2 outlines for the unique examples categorized by what EM determined to be the “correct” labels based on the Mechanical Turk labels. This set of unique examples with valence labels are used to we train a classifier Study 1.

Table 4.3.1.2 Student Examples and Associated Mechanical Turk Labels

	Positive	Negative	Neutral	Mixed	Total
2016	520 33%	398 25%	679 29%	181 13%	1778

4.3.1.3 2017C DATA COLLECTION (TEST DATA)

In the 2017C dataset 258 out of 447 (58%) students provided at least one message in the post-activity. The post activity used when generating the test data set asked students to provide examples of positive, negative, neutral, and mixed with the following instructions:

For the next set of questions, we ask you to look at all the messages in the conversation log in Udio. When looking at all the messages, try to identify which message contains a **positive** or **negative** reaction to the data from the World Bank. Not all messages will easily fall into positive or negative reactions, however. Some messages may lack emotion and can be considered **neutral**. Some messages may contain both positive and negative content and should be considered **mixed**. Here are some examples of sentences that are positive, negative, neutral, and mixed, and ambiguous:

- **Positive:** “I really like that the Netherlands contributes 5.5% of their GDP for education”
- **Negative:** “The Netherlands only contributes 5.5% of their GDP for education which is not enough”
- **Neutral:** “The Netherlands contributes 5.5% of their GDP for education”
- **Mixed:** “It is good that the Netherlands contributes 5.5% of their GDP for education, but unfortunately they do not spend the money wisely”

The 1008 examples were comprised of 755 unique messages. These unique messages were categorized by the EM algorithm which resulted in 196 positive messages, 127 negative messages, 349 neutral messages, and 83 mixed messages (see Table 4.3.1.3).

Table 4.3.1.3 Ground Truth Labels of chat messages into four emotion categories in 2017C

	Positive	Negative	Neutral	Mixed	Total
2017C	196 (26%)	127 (17%)	349 (46%)	83 (11%)	755 (100%)

4.3.2 RQ1: TO WHAT EXTENT DO STUDENTS AGREE IN TERMS OF INTER-RATER AGREEMENT WHEN PROVIDING EXAMPLES AS COMPARED WITH MECHANICAL TURK RATERS?

4.3.2.1 RQ1A: TO WHAT EXTENT DO STUDENTS AGREE IN TERMS OF INTER-RATER AGREEMENT WHEN PROVIDING EXAMPLES?

For the 2017 data set I consider examples provided by 447 students. This resulted in 1005 records where a student provided a message as examples of the aforementioned four of positive, negative, neutral, and mixed. In the 1005 records there were 755 distinct messages. No students provided the same example in more than one valence category. I computed a Fleiss' Kappa of 0.41 for all messages with two ratings, 0.50 for all messages with three ratings, 0.36 for all messages with four ratings, and -0.15 for all messages with five ratings. Finally, I computed Krippendorff's alpha statistic of 0.42 for all messages. (see Table 4.3.2). For both data sets (2016 and 2017C) Krippendorff's alpha was within the range of moderate agreement 0.40 and 0.60 previously used by researchers considering the valence categories of positive, negative, neutral and mixed (Schmidt & Burghardt, 2018) and better than studies which examined agreement on highly subjective topics which produced Krippendorff's alpha of 0.01 (Alonso et al., 2013). However, it was still below the mean agreement statistic of 0.60 for social computing (Salminen et al., 2018). These agreement statistics indicate that the level of reliability comparable with valence considering four categories. To answer the question: RQ1A: "To what extent do students agree in terms of inter-rater agreement when providing examples?" I consider the agreement of student labels for 2016 and 2017C to be sufficiently reliable as they are comparable with reliability results of published

research with the caveat there is room for improvement to align with social science research.

Table 4.3.2.1 Student Sourced Labels Agreement Statistics

Data	1 rating	2 ratings	3 ratings	4 ratings	5 ratings	Krippendorff's alpha
2016 Student Labels (1979)	1586 (-)	330 (0.42)	56 (0.52)	6 (0.30)	1 (-)	0.44
2017C Student Labels (755)	577 (-)	139 (0.41)	30 (0.50)	4 (0.36)	5 (-0.15)	0.42

4.3.2.2 RQ1B: TO WHAT EXTENT DO MECHANICAL TURK RATERS AGREE IN TERMS OF INTER-RATER AGREEMENT WHEN PROVIDING LABELS FOR STUDENT SOURCED EXAMPLES?

I asked Mechanical Turk raters to rate the 1778 messages categorized as positive, negative, neutral, or mixed by the EM algorithm from student labels (see section 4.3.1.1). Mechanical Turk workers were asked to label the messages as positive, negative, neutral, or mixed. With five workers rating each message this generated 8890 labels. We then computed Krippendorff's alpha of 0.25 and I interpret this to be fair agreement (Landis & Koch, 1977) which is below the threshold I would consider effective for the task. The results indicate agreement was closer to a study of four valence categories which produced agreement of 0.22 (Schmidt & Burghardt, 2018) where the researchers chose to not use the four categories and opted instead to use ratings of two categories which had moderate agreement of 0.47. To answer the RQ1B: "To what extent do Mechanical Turk raters agree in terms of inter-rater agreement when providing labels for student sourced examples?", we found insufficient agreement. While this suggests the data is not usable this set of labels were collected to compare with the student labels (RQ1A) and we use the data collection to examine RQ2 as a comparative benchmark for classifier training.

There are a variety of strategies which could be used to increase the inter-rater reliability of crowd sourced labels. We could revise instructions to increase agreement or examine individual rater quality and exclude rater. Neither of these approaches are explored as the intent of this data collection is to examine five raters outside of the context to the student labels using the same instructions and report the comparative analysis. As none of the students are exclude we also do not exclude any MTurk raters.

Table 4.3.2.2 Student Sourced Labels Agreement Statistics

Data	5 ratings	Krippendorff's alpha
2016 Mechanical Turk Labels (1778)	1778 (0.25)	0.25

4.3.2.3 ANSWERING RQ1: TO WHAT EXTENT DO STUDENTS AGREE IN TERMS OF INTER-RATER AGREEMENT WHEN PROVIDING EXAMPLES AS COMPARED WITH MECHANICAL TURK RATERS?

Mechanical Turk raters generating valence labels (positive, negative, neutral, and mixed) for 1778 messages resulted in a Krippendorff's alpha of 0.25 which was lower than agreement by student labels using five categories (positive, negative, neutral, mixed, and ambiguous) Krippendorff's alpha = 0.44 as well as student labels using the same four categories (positive, negative, neutral, and mixed) Krippendorff's alpha = 0.42. We consider these results to indicate student sourcing is usable as there is moderate agreement (Landis & Koch, 1977). This decision is further supported as students had better agreement when compared with Mechanical Turk ratings (0.25). To answer RQ1: "To what extent do students agree in terms of inter-rater agreement when providing examples as compared with Mechanical Turk raters?" the results are aligned with published educational research indicating the process is sufficient with room for improvement.

4.3.3 RQ2: TO WHAT EXTENT CAN CROWD SOURCED, AND IN PARTICULAR STUDENT SOURCED, EXAMPLES TRAIN A MACHINE LEARNING CLASSIFIER TO PREDICT THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

4.3.3.1 RQ2A: TO WHAT EXTENT CAN STUDENT LABELS TRAIN A LOGISTIC CLASSIFIER WHICH PREDICTS THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

To examine the extent to which a logistic classifier can be trained on student sourced labels (SSSAC Logistic) we examined the permutations of configuration for both pre-processing and processing. The pre-processing configuration considered the number of n-grams between one and three, the binary option of removing noise words from each n-gram (Yes, No), and the binary option of weighting features using TF-IDF (Yes, No). In terms of processing we examined the order of evaluating the valence classifiers of positive, negative, and mixed which is comprised of six options of ordering. This generated 168 possible configurations. We then used the 2016 data with student labels to train the classifier and the 2017C data to test the classifier. We first report results the five best F-Scores on the test data as well as the cross-validation scores from the training data for the same five configurations. The best model trained on the 2016 data and tested on the 2017C data used mono-grams, bi-grams, and tri-grams without removing any noise words, applied the TF-IDF weights to the features and processed the valence in the order of Negative, Positive, and then Mixed. This resulted in an accuracy of 0.509, Macro F-measure of 0.462. The same configuration produced a similar accuracy of 0.496, Macro F-measure of 0.475 in cross-validation.

Table 4.3.3.1a Accuracy of classifier trained on student labels predicting novel student labels

NGram	TFIDF	Order	accuracy	macro	mixed	negative	neutral	positive
1,2,3	Yes	Neg,Pos,Mix	0.509	0.462	0.317	0.379	0.602	0.550
1,2,3	Yes	Neg,Mix,Pos	0.511	0.460	0.300	0.379	0.602	0.561
1,2*,3	Yes	Neg,Pos,Mix	0.507	0.459	0.309	0.384	0.602	0.540
1*,2,3	Yes	Neg,Mix,Pos	0.511	0.458	0.290	0.383	0.605	0.554
1,2	Yes	Neg,Pos,Mix	0.507	0.458	0.295	0.390	0.602	0.544

Note: * indicates noise words were removed prior to generating n-grams

Table 4.3.3.1b Accuracy of classifier trained on student labels using 10-fold cross validation

NGram	TFIDF	Order	accuracy	macro	mixed	negative	neutral	positive
1,2,3	Yes	Neg,Pos,Mix	0.495	0.475	0.338	0.480	0.476	0.605
1,2,3	Yes	Neg,Mix,Pos	0.501	0.476	0.329	0.480	0.476	0.620
1,2*,3	Yes	Neg,Pos,Mix	0.487	0.466	0.312	0.470	0.477	0.605
1*,2,3	Yes	Neg,Mix,Pos	0.494	0.464	0.289	0.465	0.475	0.625
1,2	Yes	Neg,Pos,Mix	0.491	0.470	0.327	0.470	0.476	0.606

Note: * indicates noise words were removed prior to generating n-grams

4.3.3.2 RQ2B: TO WHAT EXTENT CAN MECHANICAL TURK LABELS TRAIN A LOGISTIC CLASSIFIER WHICH PREDICTS THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

To examine the extent to which a logistic classifier can be trained on MTurk labels we examined the same 168 possible configurations detailed in section 4.3.3.1. We then used the 2016 data with MTurk labels to train the classifier and the 2017C student labels to test the classifier. We first report results of the five best F-Scores on the test data as well as the cross-validation scores from the training data for the same five configurations.

The best model trained on the 2016 MTurk labels and tested on the 2017C student labels used mono-grams, bi-grams, and tri-grams removing noise words from tri-grams, applied the TF-IDF weights to the features and processed the valence in the order of Negative, Positive, and then Mixed. This resulted in an accuracy of 0.517, Macro F-measure of 0.456. The same configuration produced an accuracy of 0.584, Macro F-measure of 0.550 in cross-validation. The higher scores in cross validation indicate over-fitting the training data which in this case indicates over-fitting the labels provided by Mechanical Turk ratings.

Table 4.3.3.2a Accuracy of classifier trained on mechanical turk labels predicting novel student labels

NGram	TFIDF	Order	accuracy	macro	mixed	negative	neutral	positive
1,2,3*	Yes	Neg,Pos,Mix	0.517	0.456	0.341	0.378	0.623	0.480

2	Yes	Neg,Pos,Mix	0.519	0.455	0.329	0.385	0.627	0.480
3	Yes	Neg,Pos,Mix	0.514	0.454	0.342	0.375	0.622	0.477
2	Yes	Neg,Mix,Pos	0.519	0.453	0.321	0.385	0.627	0.481
1,2,3*	Yes	Neg,Mix,Pos	0.515	0.451	0.323	0.378	0.623	0.480

Note: * indicates noise words were removed prior to generating n-grams

Table 4.3.3.2b Accuracy of classifier trained on Mechanical Turk labels using 10-fold cross validation

NGram	TFIDF	Order	accuracy	macro	mixed	negative	neutral	positive
1,2,3*	Yes	Neg,Pos,Mix	0.584	0.550	0.408	0.499	0.637	0.654
2	Yes	Neg,Pos,Mix	0.589	0.553	0.410	0.496	0.647	0.660
3	Yes	Neg,Pos,Mix	0.585	0.553	0.424	0.505	0.635	0.648
2	Yes	Neg,Mix,Pos	0.592	0.548	0.373	0.496	0.647	0.675
1,2,3*	Yes	Neg,Mix,Pos	0.588	0.546	0.379	0.499	0.637	0.671

Note: * indicates noise words were removed prior to generating n-grams

4.3.3.3 RQ2C: HOW DO LOGISTICS CLASSIFIERS TRAINED USING STUDENT LABELS AND MECHANICAL TURK LABELS COMPARE TO GENERAL BENCHMARKS WHEN PREDICTING THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

In order to address research question 2, I calculated the f-measures for eight lexical approaches: VADER, SentiStrength, GI, Hu-Liu, EmoLex, Lasswell, LIWC and GALC; and two heuristic approaches: Random Baseline (RB), and Majority Class Baseline (MCB); and finally, I used the 1,778 messages to the machine learning classifiers: SSSAC Logistic. Using ten-fold cross validation I computed the f-measures for SSSAC Logistic.

As illustrated in Table 4.3.3.3, the best classifier when evaluating on novel data in overall accuracy was SSSAC Logistic (f-measure = 0.462) as compared with a Logistic classifier trained on Mechanical Turk labels (f-measure = 0.456). Both specialized classifiers outperformed all general measures and heuristics which was expected based

on our literature review classifiers trained in a specific domain. While the classifier trained on student labels had a higher level of accuracy in terms of cross-validation (f-score 0.550) than the classifier trained on student labels (f-score = 0.475) the classifier trained on student labels had the highest accuracy score when evaluated on new data (f-score 0.462) labeled by students in the 2017C dataset. The larger drop between cross-validation and testing on novel data for Mechanical Turk generated labels suggests that the model trained on Mechanical Turk labels overfit the training data which means there were features in the training data that were noise that did not help train the model to predict the test data set 2017C.

Table 4.3.3.3 Accuracy of Predicting Student Sourced Labels with F-Measures

Method		2016 (1778)	2017C (755) (using student labels)
Machine Learning Classifier	Logistic Regression	0.475† (using student labels)	0.462††
	Logistic Regression	0.550† (using MTurk labels)	0.456††
Lexical (Bivariate Mixed)	VADER	0.43	0.43
	SentiStrength	0.41	0.38
	LIWC	0.40	0.39
	GI	0.40	0.28
	Hu-Liu	0.39	0.35
	EmoLex	0.34	0.29
	GALC	0.30	0.27
	Lasswell	0.29	0.25
Heuristic	RB	0.24	0.25
	MCB	0.12	0.16

Note. † 10-fold Cross Validation Using $F_{tp,fp}$
†† trained on 2016 data to predict 2017C data

4.3.3.4 RQ2D: TO WHAT EXTENT DO STUDENTS FIND PREDICTIONS FROM A STUDENT SOURCED CLASSIFIER USEFUL?

Across the six students interviewed they reviewed 113 messages of which they agreed with the algorithm 36 times, and disagreed 77 times. For the 77 disagreements they changed their mind to agree 21 times (27% or 21/77) after seeing the algorithm's prediction. When considering the initial agreement (36 times) and when they changed their mind (21 times) the students considered the prediction accurate 50% of the time (57/113). I finally asked if sentiment analysis was useful when making predictions of their messages. Interviewees wrote down a response to this question. I coded the written response as yes or no and found five out of six students interviewed (83%) said that it was useful (see Table 4.3.3.2).

Table 4.3.3.2 Agreement, Disagreement, Final Agreement, and Usefulness of SSAC

Student	Agree	Disagree (change)	Final Agree %	Useful	Is it useful?
1	7	10 (1)	47%	Yes	I think that the help of the computer for analyzing my comments helped me understand the meaning behind my comments better and will also be of help in the future, when the system is made better (e.g. anger management, etc.)
2	9	21 (11)	67%	Yes	Really encourage that. It can help me to better understand the answers by others.
3	8	9 (2)	59%	Yes	Yes, I think it was really useful. Being able to compare my results to the computers result, helped me to see how artificial intelligence improved.

4	5	13 (3)	44%	Yes	Yes. How my message pass across.
5	6	13 (2)	42%	Yes	I think it was useful in showing me an alternative argument. There were more than one case that changed my mind which shows that I learned from it.
6	1	11 (2)	25%	No	It was interesting because the technology has reasons that do make sense to get to an answer. However, I am not sure what the technology would be used for.

For the most part participants changed their mind to agree with the algorithm 1-3 times with the exception of one student who changed their mind 11 times. The students who found the algorithm to be useful had initial agreements (5-9 messages) and final agreements (8-20 messages) with a final agreement percent that ranged from 42% to 67%. The one student who did not find the algorithm to be useful, Student-6, only initially agreed with the algorithm once and changed their mind to agree with it two times for a total of three agreements out of twelve messages (25%). Participants who found the algorithm useful agreed with the algorithm more ($M=0.52$; $SD=0.10$) than the participant who did not find it useful who agreed only 25% of the time, $t(4) = 5.712$, $p < .001$.

When describing the usefulness of the algorithm. Participants described benefits including: 1) better understanding their own communication (e.g., “I think that the help of the computer for analyzing my comments helped me understand the meaning behind my comments better...”), 2) better understanding communication of other students (e.g. “Really encourage that. It can help me to better understand the answers by others.”), and 3) seeing an alternate interpretation that changed their mind which they described as learning from the algorithm (e.g., “I think it was useful in showing me an alternative argument. There was more than one case that changed my mind which shows that I learned from it.”).

4.3.3.5 ANSWERING RQ2 TO WHAT EXTENT CAN CROWD SOURCED, AND IN PARTICULAR STUDENT SOURCED, EXAMPLES TRAIN A MACHINE LEARNING CLASSIFIER?

The evidence supports the claim that student opinions are useful when training a classifier to predict the opinions of a new cohort of students. When six students from the new cohort were interviewed five out of six students considered the classifier trained on student opinions to be useful. The one student who did not find it useful considered the classifier to have reasons to make the predictions that it made and even changed their mind to agree with the classifier two times. However, they were not sure how the technology would be used. This result is somewhat surprising because a typical crowdsourcing method is to use Mechanical Turk to generate labels to train a classifier. MTUkr labels had higher inter-rater reliability and trained a classifier with higher cross-validation scores in conjunction with lower accuracy on predicting student labels from novel data. The overall accuracy of the student sourced classifier produced an F-score of 0.461, indicating there is a lot room for improving the approach.

The students found the student sourced classifier useful and every student interviewed changed their mind to agree with the classifier at least once. Ultimately the most significant evidence that student examples from 2016 can be used to train a classifier is that they performed better in overall accuracy than a classifier trained on MTurk labels when evaluated against student examples from 2017.

4.3.4 SAMPLE CONVERSATION, STUDENT LABELS, AND PREDICTIONS BY SSSAC LOGISTIC AND SENTISTRENGTH

In sections 4.3.2 I found SSSAC Logistic had the highest overall accuracy compared with 12 benchmark technologies (see Table 4.3.2). In section 4.3.3 I saw that SSSAC Logistic was reasonable in terms of positive, neutral, and mixed detection as it was essentially in the middle of the benchmarks. Section 4.3.3 also illustrated that the median score of recall for negative valence was lower than SSSAC Logistic for all twelve benchmarks and the difference between recall scores was significantly difference from zero for six of the twelve benchmark classifiers. To help contextualize these results this section inspected a single conversation in depth providing both context for the

results and starting to establish face validity of the measure (validity is emphasized in Chapter 6 so this section serves primarily to contextualize results from RQ1 & RQ2).

I selected the conversation with the largest number of comments that had a student sourced label. The result of this selection was a conversation between six students that had a total of 100 comments, where 20 of those comments had a student sourced label. In this section I provide an illustration of how SentiStrength and SSSAC Logistic made their valence prediction as well as provide the transcript of the entire conversation to provide context of the lab exercise. Table 4.3.4 has the 100 comments from the conversation in the order the comments were written during the lab exercise. The ‘Student Label’ column displays the student label for the message when available. When a student label was provided I also added the classifications that were made by both SentiStrength and SSSAC Logistic. In RQ1 and RQ2 I used 1778 student sourced labels. This conversation has 20 of the 1778 (1.1%). In this conversation the student sourced labels were 5 positives (25%), 4 negatives (20%), 7 neutral (35%), and 4 mixed (20%). This composition was slightly different from the overall sample, which was 33% positive, 25% negative, 29% neutral, and 13% mixed (see Table 4.3.1b). The intent of this inspection is simply to gain some context and provide an early indicator of face validity (for a systematic focus on validity see Chapter 6). Table 4.3.4 illustrates this conversation.

Table 4.3.4a - An Example Conversation

User	ID	Content	Student Label	SentiStrength	SSSAC Logistic
Student_01	1	hi			
Student_02	2	Hi , I am [Student_02] and I am from [Country_01] the [Nationality_01] part>			
Student_03	3	Hello everyone, as you probably know my name is [Student_03], [#] years old, from the [Country_02] and I have the [Nationality_02] and [Nationality_03] nationality.			
Student_04	4	Hi, I am [Student_04], I am [#] years old and I am from [Country_03] but partly [Nationality_02] as well.			
Student_01	5	hi my name is [Student_01] I am from [Country_04] and I like pizza	0	+	+
Student_05	6	hello, I am [Student_05] and I am from [Country_05]			
Student_06	7	Hi, I am [Student_06] from [Country_03].			
Student_03	8	so guys, we need to discuss now for 30 minutes....			
Student_01	9	So our countries are [Country_05], [Country_02], [Country_01], [Country_03] and [Country_04]			
Student_03	10	yes			
Student_02	11	Yes, maybe we can start to compare with the different Gender !!!			

Student_01	12	I have checked the report from [Country_04] and there were [#] children out of studies, so what about your countries?			
Student_04	13	In [Country_03] it is [#] children			
Student_06	14	[Country_03] has [#] out of school children			
Student_02	15	In [Country_01] [#] children			
Student_03	16	1745 for the [Country_02]			
Student_06	17	So a low score is better?			
Student_03	18	but that does not say so much since there is a difference in total population			
Student_04	19	Maybe we first discuss the numbers of the tabel and then we will go on with discussing which country needs extra funding to make a conclusion. First of all its important that we know the numbers of the other countries.			
Student_02	20	Yes so for now is [Country_04] !!!	+	+	+
Student_04	21	alright! [Student_01] do you wanna tell us something about the numbers you have and which ones stands out?			
Student_01	22	net enrollment for pre primary school is [#]% for both sexes, [#]% for primary school and [#] for the secondary school			
Student_03	23	has anyone a "Gender parity index for gross enrolment ratio. Primary" lower than 1?			
Student_04	24	For [Country_03] the net enrollment for pre primary and secondary isnt given. For primary it is [#] %. Do you have the same [Student_06] ?	±	0	0
Student_02	25	For [Country_01], the net enrollment for the pre primary school for both sexes [#] for primary school [#] and for secondary [#].			
Student_01	26	So compare to Europe [Country_04] has lower percentage in primary school	0	0	0
Student_04	27	i would say so			

Student_04	28	[Student_05] what about [Country_05]?			
Student_03	29	for the [Country_02]; [#] [#] [#] [#]	0	0	0
Student_02	30	[Student_03] I do not have less than one for the gender parity.	±	0	0
Student_04	31	and what is the goverment expenditure on education in your country?			
Student_03	32	[#] percent	0	0	0
Student_04	33	[#] in [Country_03]	0	-	0
Student_01	34	[#]			
Student_03	35	the Gender parity index for gross enrolment ratio. Primary is for everyone the same so that is okay			
Student_02	36	[#] [Country_01]			
Student_03	37	i don't get why my primary completion rate is more than [#] percent?????	±	0	-
Student_05	38	for [Country_05] is [#]			
Student_04	39	[Student_05]??still there?	-	0	±
Student_04	40	what about [Country_05]?			
Student_04	41	[#] for what?	-	-	-
Student_06	42	it is an index not percent			
Student_01	43	[#] percent I guess			
Student_04	44	can you give us some more numbers from your country so we can take [Country_05] into account while comparing			
Student_05	45	Yes sorry. for [Country_05] is [#] the goverment expenditure on education	0	-	-
Student_06	46	lets just concentrate on the [#]-[#] most important			

Student_03	47	okay, I think that the enrollment in primary school is most important			
Student_04	48	and which are they would you say?			
Student_04	49	alright and what about out-of-school children?			
Student_05	50	the numbers are : [#] [#] [#] [#]			
Student_05	51	[#] children out of school	0	0	-
Student_03	52	next to that it is the point where our countries differ the most, [Country_04] has this percentage very low compared to the rest			
Student_01	53	Pre primary is [#]% Primary [#] Secondary [#]% Government expenditure is [#] %			
Student_06	54	primary enrollment and primary competition			
Student_04	55	okay than everybody please give the numbers for those two and their country			
Student_05	56	i think [Country_05] has the highest percentage for children out of school	-	0	0
Student_01	57	primary Competition is [#]			
Student_04	58	[Country_03]: [#] primary enrollment [#] primary completion			
Student_06	59	enrollment [#] completion [#]			
Student_01	60	Primary competition is [#] Primary enrollment is [#]			
Student_02	61	And I don't think [Country_01] need because the invest a lot.			
Student_05	62	[Country_05] [#] primary enrollment [#] primary completion			
Student_03	63	[Country_02]: [#] primary enrollment [#] primary completion			
Student_04	64	just give us the two numbers please			
Student_01	65	So guys from these data we can see that my country has the lowest percent of enrollment in a high school			

Student_06	66	i think [Country_04] has some problems with enrollment	-	0	-
Student_04	67	so what we can see is that the number of primary completion from [Student_01]s country is pretty high in comparison with the other countries. do you agree? And for primary enrollment the [Country_02] has the highest percentage.			
Student_06	68	lets choose them			
Student_01	69	So we can conclude that primary school is important and money should be invested to the primary school			
Student_03	70	me too			
Student_04	71	yes [Student_06] your right			
Student_02	72	[Country_01] Primary enrollment [#] Completion [#]			
Student_06	73	like always			
Student_05	74	ok so we can submit our answer for [Country_04]			
Student_04	75	Due to those numbers what country would need extra fund in your opinion?			
Student_06	76	[Country_06]			
Student_03	77	Let's choose [Country_04], because we think that primary enrollment is one of the most important issue and [Country_04] is the county who has the lowest percentage			
Student_01	78	Children are flowers of the life and they need good education			
Student_06	79	ok and what do we have to do now ?			
Student_02	80	I agree with [Student_03]	+	0	+
Student_02	81	Step 4 no ?			

Student_03	82	so can we proceed to step 4?
Student_06	83	ok
Student_05	84	its step four. one of our group members need to submit the answer
Student_04	85	yeah [Student_03] your right! So we have our answer..
Student_01	86	Yes go on and submit our
Student_03	87	guys do we agree on [Country_04]??
Student_03	88	someone can write that tha as our answer
Student_02	89	Yes
Student_06	90	yes
Student_03	91	i can write it now
Student_01	92	Our answer is [Country_04]
Student_03	93	Our answer is [Country_04]
Student_04	94	just one
Student_03	95	Nice [Student_01] ;-)
Student_03	96	i'm gonna do the survey in step 5
Student_04	97	okay me too
Student_01	98	Our answer is [Country_04] , because we decided that primary school is the most important part of school where [Country_04] has the lowest percentage compare to other countries
Student_03	99	where do you live???
Student_06	100	?

+	+	+
+	0	+
+	+	+
±	-	0

Key

+	-	0	±
Positive	Negative	Neutral	Mixed

All three labels were the same in 7 out of 20 messages (35%). An example of a statement where the students, SentiStrength, and SSSAC Logistic all considered the statement to be positive is message ID=95: “Nice [Student_01] ;-)”. An example of a neutral statement where all three labels agreed is message ID=26: “So compare to Europe [Country_04] has lower percentage in primary school”. An example of a negative statement where all three labels agreed is message ID=41: “[#] for what?”.

The most common occurrence, 10 out of 20 times (50%), was to have two out of three labels agree with each other. SSSAC Logistic and SentiStrength agreed with each other and differed from the student label 5 out of 20 times (25%). SSSAC Logistic and the Student Label agreed with each other and differed from SentiStrength 4 out of 20 times (20%). SentiStrength and the Student Label agreed with each other and differed from SSSAC Logistic 1 out of 20 times (5%). Which means that SSSAC Logistic was only the minority opinion in 1 out of 20 situations where at least two labels agree.

There are only three messages (15%) where all three labels disagree with one another. An example of a message with three different labels is message ID=37: “i don't get why my primary completion rate is more than [#] percent?????”.

4.4 DISCUSSION

The main purpose of Study 1 was to examine how a SSSAC Logistic trained on univariate mixed emotion examples generated by crowd sourcing valence labels from students evaluating their own online group discussions compared with general SA and contrasted with Mechanical Turk labels.

research question 1 focused on the reliability of student labels which produced mixed results. The two data sets examined, 2016 and 2017, resulted in Krippendorff's Alpha of 0.44 and 0.42 respectively both indicating moderate agreement (Landis & Koch, 1977). This agreement level is below the normal agreement in social science research (0.60) (Salminen et al., 2018).

research question 2 compared two machine learning algorithms (trained on student labels and MTurk labels); eight lexical approaches; and two heuristic benchmarks. The SSSAC Logistic trained on student labels was identified as the best in terms of overall accuracy indicating that training on a sample of data labeled by students resulted in a set of data capable of training a classifier that was more effective than the comparison technologies suggesting there is some merit to our sampling strategy. This result aligned with the AL literature (Shickel et al., 2016) suggesting that I use four valence categories of positive, negative, neutral, and mixed when asking students to label data as student labels produced the highest accuracy in predicting labels from a different group of students.

It may be worth noting that from the literature review I anticipated that around 10% of messages would most accurately be described as mixed (see section 2.1.7.2). Given that the student sourced process resulted in 13% of examples as mixed valence this aligns with expectations on the proportion of mixed communications. Further work would be needed to determine if the result of student sourcing examples using the method described in Study 1 produces a proportion of examples in each valence category that reflects proportion of all the communication from which the examples were selected.

Another result of student sourcing labels is that I saw examples of confusion, categorized by students as mixed valence messages. This is worth further examination as there are a variety of research methods in learning analytics and educational datamining that focus on the detection of confusion (Baker et al., 2012; D'Mello & Graesser, 2012; D'Mello, Lehman, Pekrun, & Graesser, 2014). Given that there are so many methods to detect confusion in students, this may provide a means to triangulate the detection of mixed emotional valence through multi-modal analytics that could include facial recognition of confusion in conjunction with mixed valence detection in text.

4.5 LIMITATIONS AND FUTURE RESEARCH

While Study 1 introduced a measure using crowd sourcing methods which outperformed 10 benchmark technologies, a known limitation of crowd sourcing is that

the quality of the measure is dependent on the ability of the crowd. When inspecting the example conversation (see section 4.3.4) there were multiple ways to interpret the results. Furthermore, as students were asked to select a maximum of 3 messages for each of the four valence categories, obviously a lot of messages were not coded by students, which may provide important insights into what students attend to when prompted to identify emotion in text. The most generous interpretation for SSSAC Logistic would be to consider SSSAC Logistic correct when either all labels agree (for our sample conversation this occurs 7 out of 20 times) or when two out of three labels agree and SSSAC Logistic is in the majority (for our sample conversation this occurs 9 out of 20 times). Combined this would have SSSAC Logistics accuracy around 16 out of 20 times (80%) on the sample conversation. This interpretation would suggest that when two classifiers made the same prediction their prediction was more accurate than the student provided label. Alternatively, the more conservative interpretation would be to consider SSSAC Logistic correct only when it matches the student label. This is the method used when predicting the accuracy of the classifier. When using this approach places to examine SSSAC Logistic's accuracy for the sample conversation I use the combination of messages where all three labels agree (for our sample conversation this occurs 7 out of 20 times) and the situation where SSSAC Logistic and the Student Label agree while SentiStrength had a different prediction (for our sample conversation this occurs 4 out of 20 times). Combined this would indicate SSSAC Logistic is correct 11 out of 20 times (55%) for the sample conversation. It is important to follow up Study 1 with research questions related to gaining more insight into the accuracy of the student raters. One approach is to consider the question: to what extent can SSSAC Logistic be replicated with a different crowd.

Finally, while this comparison showed SSSAC Logistic as the best predictor analyzed at predicting the valence categories students would identify it also raised questions about reliability. Further work is needed to replicate this work and explore how to improve the reliability of student labels. This leads to Study 2 where I examine how I can improve the reliability of student ratings with a randomized control trial examining the extent to which scripting discussions can impact reliability of rating.

CHAPTER 5 STUDY 2 – IMPROVING A STUDENT SOURCED SENTIMENT ANALYSIS CLASSIFIER WITH EMOTIONAL DESIGN USING EMOTIONAL SENTENCE STARTERS - A RANDOMIZED CONTROL TRIAL



5.1 INTRODUCTION

In Study 1 (Chapter 4) I found that when measuring the accuracy of sentiment analysis (SA) based on the perspectives of 767 business students that the student sourced sentiment analysis classifier (SSSAC) logistic had the best overall accuracy (see a detailed analysis in sections 4.3.2 and 4.3.3). SSSAC logistic was trained on valence labels for 1778 text messages in our online learning platform called Udio that came from crowd sourcing methods, resulting in a tested overall accuracy summarized by the F-measure of 0.462. While this score was higher than ten benchmarks and a competing crowd source classifier using Mechanical Turk, this simply indicated that the majority opinion of students (represented by the crowd sourcing EM algorithm) had more differences from ten general benchmark SA methods and the algorithm trained on Mechanical Turk labels than it did from SSSAC logistic trained on student labels. One key challenge is the inability for students to agree on the valence label of messages. To

build on the results from Study 1, Study 2 explores how to increase student emotion awareness to in turn improve the outcomes of student sourcing an SA classifier.

When supporting the students to express their opinions Study 2 builds on the work of Universal Design for Learning (UDL), which is a framework of design guidelines to support the development of expert learners (see section 2.1.2.3 for a more comprehensive description of UDL). According to the UDL Guidelines criticisms should fall on the environment before they fall on the student. In the context of Study 2, when asking whether or not students are effective at identifying emotion expression in text, the criticism is on the online environment and the limitations of the environment, before placing any criticism on the students. For example, when considering how to support the emotional link to learning, a UDL recommendation is to use contemporary tools (Posey, 2018) as suggested by UDL guideline 5.2 (CAST, 2018), with a specific recommendation to use sentence starters (Posey, 2018), which are a form of scripting designed to supporting text communication. In addition to the UDL direct guidance of using sentence starters to support emotion communication, the Computer Supported Collaborative Learning (CSCL) community proposes that sentence starters might be useful when explicitly supporting reflecting during the phases of self and socially regulated learning (Järvelä & Hadwin, 2013). This aligns with the perspective that students may not be perfect historians of their own emotions (Pham, 2004), because by asking participants to explicitly state their emotional reactions in text it captures a written record of how they felt at the time, and can be referenced later when they might no longer recall the emotional experiences. Sentence starters have the potential to support both the expression and reflection of emotion in chat communication. For a detailed discussion of the conceptual framing of sentence starters, I refer to section 2.2.2.3.

When referring to student sourcing labels I ask students to identify text messages in Udio as positive, negative, neutral, and mixed. In part these categories are ideal for students to reflect on their and their peers' emotions expressed in the collaborative learning tool in Udio, and with the crowd sourcing method of students reflecting on their own discussions may benefit from sentence starters. In this Study 2, I will focus on scaffolding that asks participants about their emotional reactions to the learning material

(i.e., emotional sentence starters). By introducing emotional sentence starters (ESS), I anticipate that this may improve the ability for students to recall their own emotions and potentially support students identifying the emotions explicitly written down by their peers (For a more in-depth discussion about potential strengths and weaknesses of ESS see section 2.4.2.2). The introduction of ESS raises research question (RQ3): To what extent can emotional sentence starters improve the inter-rater reliability of student examples?

As the result of conducting study 2, new example data were generated that I can use to train SSSAC Logistic. This generates new possibilities to examine how the quality and quantity of training data might improve the accuracy of the classifier. Considering the quantity and quality of training data this raises research question 4 (RQ4): To what extent can emotional sentence starters generate student examples capable of training a more accurate classifier which predicts the valence categories of positive, negative, neutral, and mixed?

In summary, the first and primary purpose of Study 2 is to see if emotional sentence starters increased the reliability of student ratings (RQ3). The second purpose is to test the capacity for unscripted examples from the ESS intervention improves SSSAC Logistic considering the quantity and quality of data (RQ4).

5.2 METHODS

5.2.1 SETTING

The study took place at a University in the Netherlands in a freshman statistics course. The University recruited international students as part of the teaching philosophy at the University was that students could learn from a diverse group of peers. In this context students typically had a problem-based learning (PBL) curriculum meaning that they were used to working in groups to solve a specific problem, see Tempelaar et al. (2015, 2017) for a detailed description of the educational context in which Study 1 collected the data. As described in greater detail in Mittelmeier et al. (2018), the World-Bank assignment asked students to discuss data from a set of countries and work on a problem of making a funding decision in a group. The group

nature of working on a problem together made this assignment ideal for this context. The primary distinction between the assignment and common experience of students was the use of an online platform, Udio, to facilitate the group work via an online chat interface.

5.2.2 PROCEDURE

In this replication study at the same business program at the same university in the Netherlands in week 6 of the respective course as in Study 1, 884 students participated in a 60-minute lab in groups of 4 ($M=3.49$; $SD=0.89$) in 2017. In 2017, the study design examined how an intervention in a randomized control trial affected the accuracy of a SSSAC. In the control condition, the activity was similar to the 2016 Pilot Experiment with four intentional procedural changes to the lab activity.

The first change was that the initial icebreaker activity to familiarize students with Udio and the respective group they were working in during the lab study was replaced with a brief statistics sampling activity. This was specifically included as requested by the teacher of the respective mathematics and statistics module, as students in 2016 at times could not always make a link between the statistics course and the World Bank lab activity.

The second change in comparison to Study 1 was that students were instructed to self-report their emotion and view the emotional reactions of their peers using the React tool in Udio. This change was designed to get a specific emotional measure of incidental emotions that students felt at the beginning of the group exercise. Figure 5.2.2a visually illustrates how students could self-report their emotional reactions using a multi-select of twelve emotion words. React provided the following list of words as options: engaging, interesting, challenging, curious, calming, good, dull, boring, sad, confusing, frustrating, and annoying. In Figure 5.2.2a the words engaging, interesting, challenging, good, boring, and sad have been selected as an example of what the student would see after selecting.

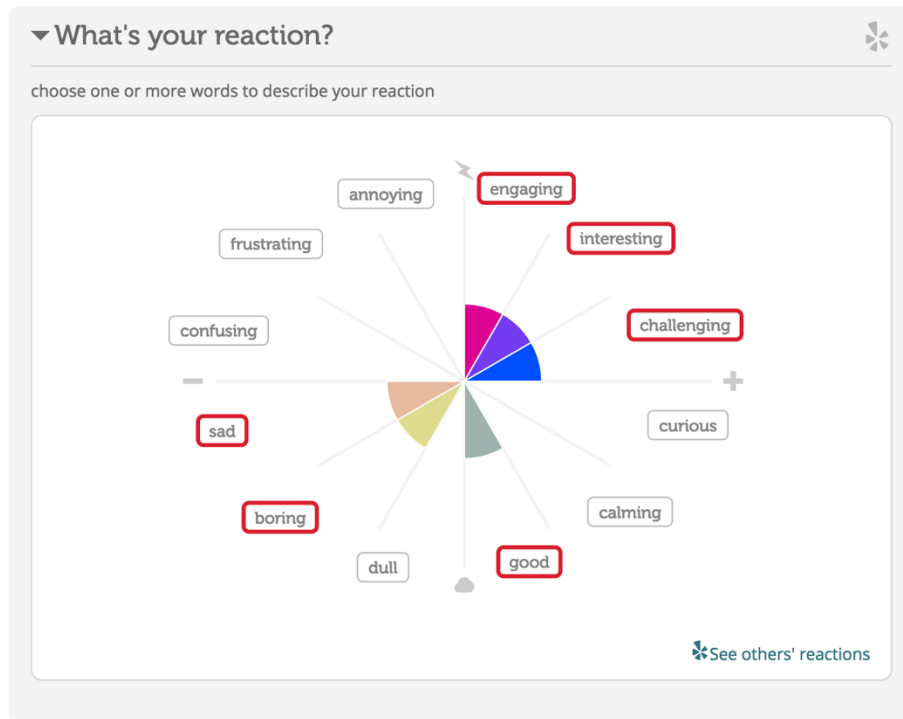


Figure 5.2.2a - Interface of 'React' to self-report emotional response

The third change was that students were assigned into a randomized condition, using the random function in excel to assign groups to students based on the students assigned to the lab section. In the control condition, students discussed the World Bank data following the case study from the pilot experiment in Study 1. In the intervention condition, in addition to work on the World Bank case students were asked to use a support during their group discussions. The support in the intervention condition provided written instruction in the main lab activity that during the group discussion the students should at least twice select from the following four sentence starters “I had a positive | negative | neutral | mixed reaction to...”. Figure 5.2.2b visually illustrates how students could select a respective sentence starter from a drop-down list.

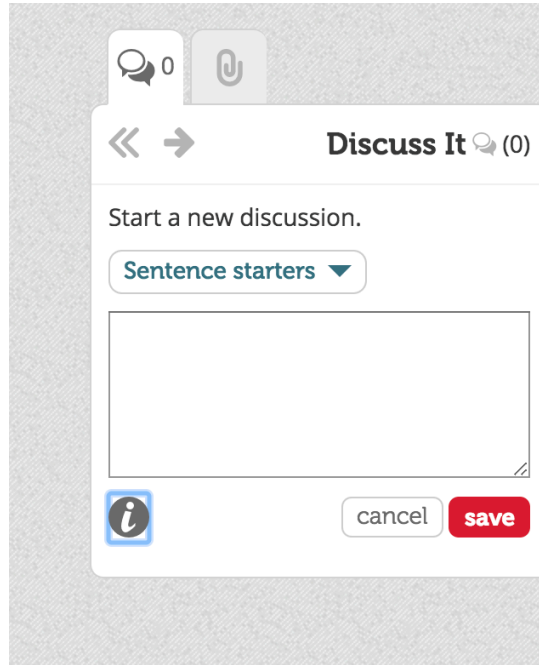


Figure 5.2.2b - Discussion interface with Sentence Starters

The fourth and final change to the process was to remove the group response component and move the answer to the case study into the output of the experiment by having students provide individual answers during an exit survey. A detailed breakdown of the Procedure is provided in Figure 5.2.2c.

Input	Process			Output			
<u>Disposition</u> BEQ	Sampling Activity (10Min)	<u>Self- Report</u> React (5 min)	<u>CONTROL</u> Discuss World Bank Data (30min) <u>INTERVENTION</u> Discuss w/ Emotion Sentence Starters (30min)	Provide an Answer (10min)	<u>Exit Survey</u> MES PANAS (10min)	<u>Post</u> Group Peer	<u>Interview</u> Accuracy Aggregate Evaluation
Before Lab	During Lab					After Lab	

Figure 5.2.2c Input, Process, Output for Main Study conducted in 2017

5.2.3 PARTICIPANTS

In Study 2 I worked with a course which had 1,075 enrolled students who were offered the opportunity to participate in a study as an alternative to doing the regularly

scheduled lab exercise. 93 Students (8.7%) chose not to participate. Another 98 (9.1%) that chose to participate had to be excluded from the analysis due to technical issues. The technical issues were related to slow response from the website hosted at the OU from the site in the Netherlands. The connectivity delay resulted in letting the affected students leave with full credit for participation. For the remaining 884 students (82.2%) they were randomly assigned to a condition through random assignment to groups where half of the groups were in the intervention condition and half were in the control condition. In the control condition 447 students participated in a 60-minute lab in groups of 4 ($M=3.49$; $SD=0.89$). In the Emotion Sentence Starter Condition 437 students participated in a 60-minute lab in groups of 4 ($M=3.49$; $SD=0.89$).

Figure 5.2.3 illustrates the assignment of students in the control and intervention conditions as well as the participants response rates based on the IPO model illustrated in Figure 5.2.2c.

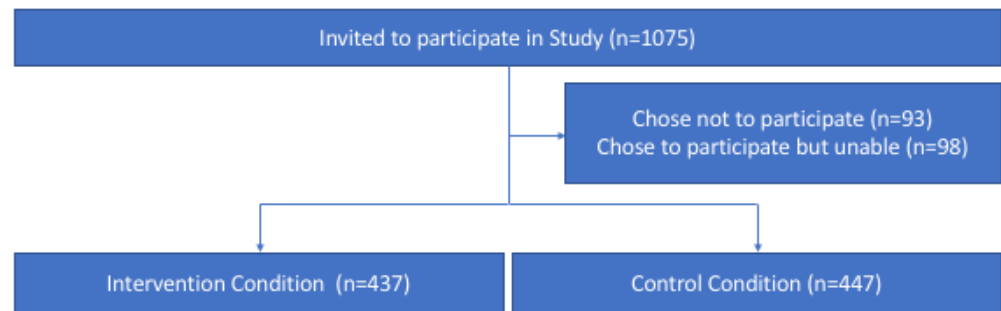


Figure 5.2.3 - Participation by Research Condition for Main Study conducted in 2017

In order to check whether there were any demographic differences between the two conditions in age, gender, and nationality, I used a range of t-tests and chi-square analyses. There was no statistical group difference in terms of age for the control

($M=19.01$; $SD=1.14$) and Intervention ($M=19.02$; $SD=1.10$) as indicated by a t-test: $t(881.96) = 0.18$, $p > .05$.

Table 5.2.3a Age by Research Condition for Main Study conducted in 2017

	Intervention N=437		Control N=447		t-test
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age	19.02	1.10	19.01	1.14	ns

Note. *M* = Mean. *SD* = Standard Deviation. ns = no significant difference.

For the control condition, the 447 students were comprised of 178 females, 262 males, and 7 unknown gender. Similarly, in the intervention condition 437 students were comprised of 189 females, 244 males, and 4 unknown gender. While gender is generally believed to have an effect on the detection of emotion in face to face interactions (Collignon et al., 2010), there is evidence that in online interactions there is no gender effect on the ability to identify emotion in text messages (Hancock, Landrigan, & Silver, 2007). There was no statistical group difference in the number of people in the gender categories of male, female, and unknown as indicated by a chi-square test ($X^2(2) = 1.68$, $p > 0.05$).

Table 5.2.3b Gender by Research Condition for Main Study conducted in 2017

Gender	Intervention N=437	Control N=447
Male	244 (250.14) [0.15]	262 (255.86) [0.15]
Female	189 (181.42) [0.32]	178 (185.58) [0.31]
Unknown	4 (5.44) [0.38]	7 (5.56) [0.37]

In the control condition the population was international with students from 38 countries around the world, including 121 domestic, 296 European Students, 21 non-European

students and 9 students for whom there was no nationality information available. In the intervention condition the population was also international with students from 33 countries around the world, including 104 domestic, 295 European Students, 30 non-European students, and 8 students for whom there was no nationality information available. There was no statistical group difference in the number of people in the nationality categories of Domestic, European, Non-European, and unknown as indicated by a chi-square test ($\chi^2(3) = 2.82, p > 0.05$).

Table 5.2.3c Nationality by Research Condition for Main Study conducted in 2017

	Intervention N=437	Control N=447
Domestic	104 (111.23) [0.47]	121 (113.77) [0.46]
European	295 (292.16) [0.03]	296 (298.84) [0.03]
Non-EU	30 (25.21) [0.91]	21 (25.79) [0.89]
Unknown	8 (8.40) [0.02]	9 (8.60) [0.02]

The final baseline measures I compared was the responses from the Berkley Expressivity Questionnaire (BEQ). Each group was asked to answer the BEQ ahead of the lab activity as a pre-survey making it a baseline measure. In the control condition 435 out of 447 students (97.32%) completed the BEQ ahead of the experiment. In the Intervention condition 420 out of 437 students (96.11%) completed the BEQ ahead of the experiment. There was no significant difference between response rates as indicated by chi-squared test for independence ($\chi^2(1) = 0.017, p > 0.05$).

The BEQ has three subscales. The first subscale is positivity which indicates the comfort level with expressing positive emotion. The second subscale is negativity which indicates the comfort level with expressing negative emotion. The third subscale is impulsivity control which indicates the extent to which the individual is capable of managing their impulsive responses. There were no significant differences for all three of the subscales as indicated by running t-tests. There was no statistically significant difference between the positivity subscale scores for intervention ($M=5.10$; $SD=1.18$) and control ($M=5.22$; $SD=1.17$) as indicated by a t-test: $t(853) = 1.493, p > .05$. There was

no statistically significant difference between the negativity subscale scores for intervention ($M=3.51$; $SD=1.24$) and control ($M=3.62$; $SD=1.32$) as indicated by a t-test: $t(853) = 1.26$, $p > .05$. There was no statistically significant difference between the impulsivity subscale scores for intervention ($M=4.45$; $SD=1.33$) and control ($M=4.55$; $SD=1.38$) as indicated by a t-test: $t(853) = 1.08$, $p > .05$. Table 5.2.3d displays the mean and standard deviation for the three subscales of the BEQ for the Intervention and Control condition.

Table 5.2.3d Berkley Expressivity Questionnaire (BEQ) scores by Research Condition for Main Study conducted in 2017

	Intervention N=420		Control N=435		t-test
	M	SD	M	SD	
Positivity	5.10	1.18	5.22	1.17	ns
Negativity	3.51	1.24	3.62	1.32	ns
Impulsivity Control	4.45	1.33	4.55	1.38	ns

Note. M = Mean. SD = Standard Deviation. ns = no significant difference.

I used a randomized control trial with the intent to get two comparable populations. As 98 participants that chose to participate were excluded, I compared baseline measures as a means of identifying if the exclusion caused a problem with the randomization. Based on the comparisons of age, gender, nationality, and the Berkley Expressivity Questionnaire (BEQ) there is no indication that the exclusion caused a significant difference in randomization. It is considered poor practice to use statistical significance testing on baseline factors to make the claim that populations are comparable (de Boer, Waterlander, Kuijper, Steenhuis, & Twisk, 2015). I simply claim that in Study 2 there is no evidence that exclusion caused problems with randomization based on the comparison of baseline measures. The two populations of the intervention and control condition are comparable in so far as randomization generates comparable populations.

5.2.4 INSTRUMENTS

5.2.4.1 STUDENT SOURCED EXAMPLES

I used the same student sourced method describe in section 4.2.4.1

5.2.4.2 STUDENT SOURCED SENTIMENT ANALYSIS INSTRUMENTS

In Study 2 I will refer to three different datasets 2016 data collected and analyzed in Study 1, 2017C data collected and analyzed in Study 1, and 2017SS new data we consider for Study 2 where is an experimental condition where students use Emotional Sentence Starters (ESS). 2017C is actually a control condition as students in 2017 were randomly assigned to either 2017C or 2017SS. Given that 2017C is analyzed in both Study 1 and Study 2 this connects the studies by building on findings from Study 1 and using 2017C as the test data set.

5.2.5 ANALYSIS

In general, I have followed the same procedures as described in section 4.2.5. I first pre-processed the text data, as indicated in 4.2.5.2. Afterwards, I processed the text data using Naïve Bayes (NB), and Support Vector Machines (SVM). For this evaluation I will use all of these two baselines approaches, SVM and NB, as machine learning benchmarks to compare with the proposed bag-of-words ensemble valence (SSSAC Logistic) classifier. Like discussed before in section 4.2.5.3, I conducted a ten-fold cross validation, to check classifier accuracy. I shifted the analysis to focus on comparing SSSAC Logistic generated in Study 1 by examining the same test data set from Study 1 (2017C) and introduce analysis using new training data (2017SS) detailed in 5.2.5.1.

5.2.5.1 COMPARATIVE ANALYSIS

In order to answer RQ3, “To what extent can emotional sentence starters improve the Inter-rater reliability of student examples?” I examine examples provided by students in the 2017SS dataset where students used Emotional Sentence Starters (ESS). As ESS is a form of scripted communication we split the 2017SS dataset into

two subsets: Scripted and Unscripted. By examining agreement between students on the unscripted examples we find that ESS increased the agreement compared to unscripted examples.

To answer RQ4, “To what extent Emotion Sentence Starters improve a student sourced sentiment analysis classifier?”, we use the unscripted examples to train a classifier and use the same test data set as used in Study 1: 2017C. I find that with fewer examples the resulting classifier is of comparable quality. As Study 1 trained a classifier with more examples generated in 2016 we construct a learning curve for both SSSAC Logistic trained on 2016 and 2017SS Unscripted to see if there is a difference in accuracy with comparable number of examples used to train the classifier. I find that not only do unscripted examples from the 2017SS dataset have a higher level of inter-rater reliability (see RQ3 results), but they also train a more accurate classifier. The implications of which suggest ESS are an effective emotion awareness tool and a novel contribution in their own right as well as supporting the creation of SSSAC Logistic.

5.3 RESULTS

5.3.1 STUDENT SOURCED LABELS

In the 2017 Experiment, a total of 884 students participated. There were two conditions: Control, and Intervention. The control condition was a replication of Study 1 conducted in 2016 with the previous cohort of students enrolled in the same class. In the intervention condition students were asked to use Emotion Sentence Starters at least twice during the group online discussion. The ESS asked students to start online discussion messages with “I had a positive|negative|neutral|mixed reaction to...”. The intent of using sentence starters was to investigate the influence it has on the student sourced method of generating a SA classifier described in detail in section 4.2.4.1. As the method depends on to students’ ability to label messages considering both the intent of the author and reaction of the reader by having students explicitly state their reactions during the discussion this has the potential benefit to student sourced examples for two reasons. The first reason is that people are not ideal historians of their own emotions

(Pham, 2004). The second reason the ESS may improve the results of student sourcing labels is that while students may have insights into the context of their own discussion groups they are not trained valence raters. By having group members explicitly state their emotions during group discussion this has the potential of making it easier to identify examples of positive, negative, neutral, and mixed messages when reflecting on their group discussions. To investigate the influence of ESS on Student Sourcing Labels I compare the control and intervention condition starting with the examples students select during their post-activity.

In the Control condition 258 out of 447 (58%) students provided at least one message in the post-activity. The 1008 examples were comprised of 755 unique messages. These unique messages were categorized by the EM algorithm, as was previously done in section 4.2.5.1. In the ESS intervention condition 269 out of 443 (61%) students provided at least one message in the post-activity. There was no distinct difference in terms of number of students who provided an example as indicated by a chi-squared test for independence ($\chi^2(1) = 0.832, p > 0.05$). This resulted in 1208 examples comprised of 752 unique messages. These unique messages were categorized by the EM algorithm as described in section 4.2.5.1. The resulting labels are presented in Table 5.3.1 as the 2017C, 2017S Examples generated in Study 2 and the 2016 Examples which were generated in Study 1 (see section 4.3.1). The distinction between the percentages of comments per valence category between the 2016 and 2017 study illustrated in Table 5.3.1 could be the result of the changes in the experimental design detailed in section 5.2.2.

Table 5.3.1 Ground Truth Labels of chat messages into four emotion categories in 2016 & 2017 (control and intervention condition)

	Positive	Negative	Neutral	Mixed	Total
2016	587 (33%)	444 (25%)	524 (29%)	223 (13%)	1778 (100%)
2017C	196 (26%)	127 (17%)	349 (46%)	83 (11%)	755 (100%)
2017SS	254 (34%)	132 (18%)	293 (39%)	72 (10%)	752 (100%)
Total Examples	1037 (32%)	703 (21%)	1166 (36%)	378 (12%)	3284 (100%)

5.3.2 RQ3: TO WHAT EXTENT CAN EMOTIONAL SENTENCE STARTERS IMPROVE THE INTER-RATER RELIABILITY OF STUDENT EXAMPLES?

The ratings provided by students produced a variable number of ratings per message ranging from one to five raters. To examine inter-rater reliability, we report Fleiss' kappa for messages with the same number of multiple ratings (I.e., 2, 3, 4, 5) and report the number of records with one rating. Using the entire data set we compute Krippendorff's alpha. The data we examine for inter-rater reliability is from three datasets. The 2016 dataset, the 2017C dataset, and the 2017SS dataset. The 2016 Dataset was collected in the first experiment conducted in 2016. The 2017 data are from a randomized control trial where students were assigned either to a control condition (2017C) which was a replication of the 2016 experiment or assigned to an intervention condition (2017SS) where each student was asked to use Emotional Sentence Starters (ESS) at least twice during online group chat. As ESS script explicit emotional reactions in terms of valence (e.g., I had a positive reaction to...) the scripting inevitably would increase the agreement of students when identifying examples of positive, negative, neutral, or mixed expressions. To address the scripting effect, we further break the 2017SS data into two mutually exclusive subsets: 1) Scripted, examples provided by students which contain the sentence starters, and 2) Unscripted, examples provided by students that do not contain ESS.

The 2016C data from the replication of 2016 experiment produced a very similar Krippendorff's alpha score of 0.42 which indicates the replication had similar outcomes in terms of inter-rater reliability of student examples. The 2017SS data set had a much higher Krippendorff's alpha of 0.71. To examine the extent to which the increased in agreement was a result from the ESS we further examined the 2017SS Scripted (Krippendorff's alpha of 0.88) and Unscripted (Krippendorff's alpha of 0.61) (For a complete breakdown of Fleiss' kappa and Krippendorff's alpha see Table 5.3.2). These results indicate that the student sourced examples in the 2016 study and the 2017C replication of that study produced Krippendorff's alpha which ranged from

Table 5.3.2 Student Sourced Labels Agreement Statistics

	1 rating (Fleiss)	2 ratings (Fleiss)	3 ratings (Fleiss)	4 ratings (Fleiss)	5 ratings (Fleiss)	Krippendorff's alpha
2016	1586	330	56	6	1	0.44
Total (1778)	(-)	(0.42)	(0.52)	(0.30)	(-)	
2017C	577	139	30	4	5	0.42
Total (755)	(-)	(0.41)	(0.50)	(0.36)	(-0.15)	
2017SS	501	145	79	21	5	0.71
Total (752)	(-)	(0.71)	(0.78)	(0.96)	(-0.15)	
Unscripted (558)	430 (-)	86 (0.63)	31 (0.61)	5 (0.84)	5 (-0.15)	0.61
Scripted (194)	71 (-)	59 (0.78)	48 (0.78)	16 (0.90)	0 (-)	0.88

5.3.3 RQ4: TO WHAT EXTENT CAN EMOTIONAL SENTENCE STARTERS GENERATE STUDENT EXAMPLES CAPABLE OF TRAINING A MORE ACCURATE CLASSIFIER WHICH PREDICTS THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

Table 5.3.4.3A Comparing and Combining 2016 and 2017 (cross validation)

Training Data	Cross Validation						
	Accuracy	Macro	Weighted	Positive	Negative	Neutral	Mixed
2016 (1778)	0.495	0.475	0.338	0.480	0.476	0.605	0.495
2017 SS Unscripted (558)	0.583	0.505	0.585	0.614	0.322	0.682	0.403

Table 5.3.4.3B Comparing and Combining 2016 and 2017 SS Unscripted data (testing on 2017C)

Training Data	Testing on Novel Data						
	accuracy	macro	weighted	Positive	Negative	Neutral	Mixed
2016 (1778)	0.509	0.462	0.317	0.379	0.602	0.550	0.509
2017 SSUnscripted (558)	0.519	0.441	0.520	0.529	0.290	0.653	0.293

I first report a learning curve examining the overall accuracy based on the number of records from the 2016 dataset. The first point to illustrate is that even with a minimum number of records analyzed (35) for the training data set the test score had an accuracy of 38%. This is likely due to the fact that records that are not identified as positive, negative, or mixed are subsequently classified as neutral. The test set has 755 records with the majority class of neutral consisting of 349 examples (46%). This indicates that a small sample generated many false positives for non-neutral valence classes. The second point is that the classifier starts to perform near the majority class baseline (46% accuracy) at 432 records where it reached an accuracy of 46%. The third point is that from 432 records to the entire sample of 1778 the curve has a small positive slope indicating the curve may not have yet peaked with the accuracy of 50.9% using 1778 records. This suggests that the classifier would likely reach a higher level of accuracy with more data.

I then examine the learning curve examining the overall accuracy based on the number of records used from the 2017 SS Unscripted data set. When constructing a learning curve using 2017 SS Unscripted data I took a subset of the training data into 16 segments (each representing 6.25% of the training data) and incrementally added segments together to train the classifier. I then repeated this process 5 times and took the average accuracy to plot a learning curve illustrating the anticipated accuracy based on the number of training records. I find that the unscripted examples from the 2017SS produce both a higher level of accuracy in terms of cross validation and in terms of testing on the 2017C data set. Testing the accuracy on novel data shows accuracy with minimal records reaches accuracy of 51.9% using the entire data set of 558 records

which is higher than the accuracy than a comparable subset of 576 messages from the learning curve which achieved a test accuracy of 46.9% and a higher accuracy than using all 1778 examples from the 2016 dataset as this produced an accuracy of 50.7%.

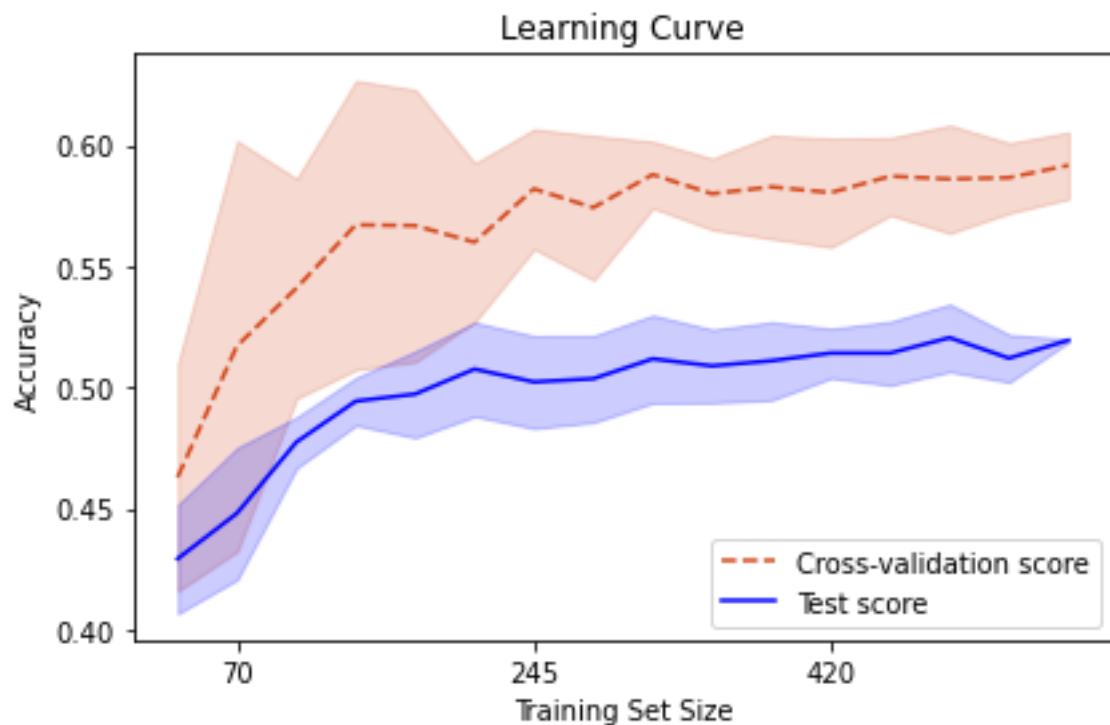
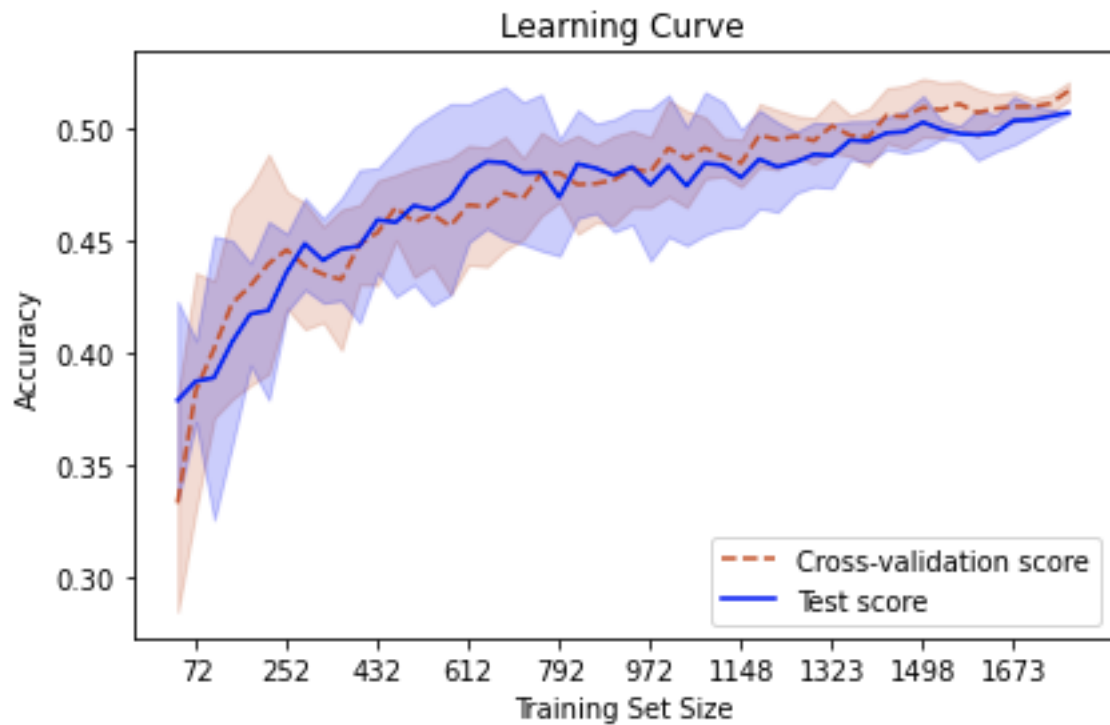


Table 5.3.4C Comparing 2017C to 2016 Subset and the entire 2016 dataset

Dataset	#	Cross	Cross	Test	
	Messages	Validation	Validation	Accuracy	Test Std
		Accuracy	Std		
2016	1778	0.495	-	0.509	-
2016 Subset	576	0.457	0.030	0.469	0.042
2017C	558	0.583	-	0.519	-

These results suggest that using the emotional sentence starter intervention produced a better classifier to detect valence using fewer records to train the classifier indicating a higher quality for the dataset suggesting increasing the quality of examples for training the classifier can produce significant improvement even with fewer high-quality examples. While it is expected that higher quality training data improves training algorithms the use of emotional sentence starters is a novel contribution to achieve that goal.

As both learning curves suggested more data would produce higher accuracy we examined combining the two training sets to see the effect on accuracy for the test set which resulted in the highest achieved f-score of 0.489 for SSSAC. See tables 5.3.4D and 5.3.4E for the results of combining the two training sets.

Table 5.3.4D Comparing and Combining 2016 and 2017 (cross validation)

Training Data	Cross Validation						
	Accuracy	Macro	Weighted	Positive	Negative	Neutral	Mixed
2016 + 2017 SS Unscripted (2336)	0.522	0.495	0.528	0.598	0.470	0.566	0.347

Table 5.3.4E Comparing and Combining 2016 and 2017 SS Unscripted data (testing on 2017C)

Training Data	Testing on Novel Data						
	accuracy	macro	weighted	Positive	Negative	Neutral	Mixed
2016 + 2017SS Unscripted (2336)	0.559	0.489	0.565	0.603	0.355	0.679	0.317

5.3.3.1 ANSWERING RQ4: TO WHAT EXTENT CAN EMOTIONAL SENTENCE STARTERS GENERATE STUDENT EXAMPLES CAPABLE OF TRAINING A MORE ACCURATE CLASSIFIER WHICH PREDICTS THE VALENCE CATEGORIES OF POSITIVE, NEGATIVE, NEUTRAL, AND MIXED?

I found that pre-processing and processing changes had nominal improvements on the SSSAC Logistic while the effect of Emotional Sentence Starters generated a higher quality set of data capable of reaching comparable accuracy with fewer records. I also found by combining the data sets the classifier reached the highest level of accuracy when testing on novel data with a macro f-score of 0.489. While increasing the number and quality of training records is not new, the approach of using emotional sentence starters to generate higher quality data is a unique contribution.

5.3.4 SAMPLE CONVERSATION FROM 2017S DATASET, STUDENT LABELS, AND PREDICTIONS BY SSSAC LOGISTIC(2016) AND SSSAC LOGISTIC(2017S)

In sections 5.3.2 I found SSSAC Logistic had comparable accuracy on a new dataset and that accuracy was improved when analyzing data where students had supports from ESS (see Table 5.3.2). Section 5.3.2 also illustrated that the mean score of recall for positive valence was higher indicating ESS make positive detection easier for the classifier trained in Study 1. In section 5.3.3 I saw that SSSAC Logistic was a replicable and that when supporting students with ESS I improved accuracy (see Table 5.3.3). Section 5.3.3 also illustrated that the mean score of recall for neutral valence was higher, indicating training with ESS improved neutral detection and overall accuracy. To help contextualize these results this section inspects a single conversation supported

with ESS in depth providing both context for the results and starting to establish face validity of the measure

I selected the conversation with the largest number of comments that have a student sourced label. For this conversation there are 14 annotations for 41 comments. 4 out of 14 student labels were predicted by both SSSAC Logistic which agreed with students 8 out of 14 times (57%).

Table 5.3.4 Sample Conversation with Labels from ESS condition

User	ID	Content	Student_Label	SSSAC Logistic
[Student_10]	01	Hi guys :)		
[Student_11]	02	Hey, I am [Student_11] and I am from [Country_01].		
[Student_12]	03	Hi, I am [Student_12] from [Country_02]		
[Student_13]	04	ehhhh how r u doin		
[Student_10]	05	I'm [Student_10] from [Country_03]		
[Student_13]	06	Im as well [Nartionality_01]		
[Student_10]	07	Maybe we can straight to the point ahah, what are the mean scores for your countries?		
[Student_13]	08	[#] [Country_01]		
[Student_10]	09	go straight*		
[Student_10]	10	[#] [Country_03]		
[Student_12]	11	[#] [Country_02]		
[Student_11]	12	[#] in [Country_01]	0	0
[Student_13]	13	what are we supposed to do now?	0	-
[Student_10]	14	we must go to step [#]		
[Student_11]	15	For [Country_01] the mean reading score is reported as [#].		

		The mean is [#], the SD is [#], the number of samples is [#]	
[Student_10]	16	Does someone has an idea for a null hypothesis?	
[Student_13]	17	I think Im doin smth wrong	- -
[Student_12]	18	H0: p equals [#] HA: p unequals [#]	
	19	Regardless what sample we are taking, the mean score for reading in [Country_03] will be [#]... Do you think it could be a correct null hypothesis?	
[Student_10]	20	I had a positive reaction to what u wrote [Student_10] because I have not a clue what the Null hypothesis is	± +
[Student_13]	21	ahah thanks Luis.. maybe this will help you http://www.statisticshowto.com/probability-and-statistics/null-hypothesis/	
[Student_10]	22	I had a positive reaction to your link, thank you	+ +
[Student_13]	23	Ok after I read the page, [Student_12]'s thing seems right	+ +
[Student_10]	24	For [Country_03]: mean: [#] sd: [#]	
	25	So what are we supposed to discuss now ? I dont get the whole point of analyzing the data we just got	- -
[Student_13]	26	Guys, to be sure, we have to choose one country among all the counties reported by PISA?	0 0
[Student_10]	27	I don't get it either, don't worry	- -
[Student_13]	28	No, I think we have to choose one of our countries	- ±
	29	we need to discuss the data of our country, did you click on the link in step [#] ?	
[Student_11]	30	Maybe we should compare the mean scores and choose the lowest one then	
[Student_10]			
[Student_10]	31	Oh right, i hadn't click on step[#]	

	32	so what can you say about the attainment over time in your country?		
[Student_11]	33	Oh boy haha, So we should divide it into our countries and then into Attainment over time, By Age Group and Gender..idk whether old ppl are interesting for us but I think we should focus on those between [#] and [#]		
[Student_13]				
[Student_12]	34	How shall we devide the work?		
[Student_12]	35	divide*		
	36	So the secondary education in [Country_01] is relatively constant since [#] and above [#] % for citizens between [#] and [#]		
[Student_13]				
	37	How do you take [#] to [#] ? I can only choose [#]-[#] then [#]-[#]..	0	-
[Student_10]	38	I think we should just talk about our own countries, I don't get the attainment over time in [Country_01], but considering by age group and gender female and male the highest percentage is secondary complete and it is about the same for male and female	±	0
[Student_11]				
[Student_13]	39	I looked up all data individually since [#] more than half of the now [#]-[#] primary education and overall less than the half have a higher educational level	0	0
	40			
[Student_12]				
	41	I had a positive reaction to the data for By Age Group and Gender. [Country_01] has a high percentage of educated citizens	+	+
[Student_13]				

5.4 DISCUSSION

In answer to RQ3, I saw a significant improvement in reliability in terms of inter-rater agreement when using examples from the 2017SS Unscripted examples. The control condition had a Krippendorff's alpha of 0.42 while the intervention condition had a Krippendorff's alpha of 0.61. This is significant as it put the agreement on par with research in social science (Salminen et al., 2018), outperformed other studies which examined the four valence categories of positive, negative, neutral, and mixed (Schmidt & Burghardt, 2018; Wilson et al., 2005), and put agreement into the category of substantial agreement which is the same category for the best agreement reviewed for studies examining positive, negative, neutral, and mixed (Chakravarthi, Muralidaran, Priyadharshini, & McCrae, 2020).

In answer to RQ4, training a classifier on unscripted examples from 2017SS had an improvement using fewer examples, 558 in 2017SS Unscripted, as compared to the baseline dataset, 1778 in 2016, to train a classifier. The most significant improvement was using examples from the intervention condition, 2017SS Unscripted, combined with examples from the 2016 dataset with a cross validated f-score of 0.495 and an f-score of 0.489 on novel data. These results suggested that more data and higher quality ratings would likely improve the classifier, which is expected. The novelty of the finding is that the emotional sentence starters played a role in producing higher quality data.

There are in fact many examples where emotional labeling is done through a crowdsourcing manner (Hutto & Gilbert, 2014; Morris & McDuff, 2009; Warriner et al., 2013). Some critics have pointed out that crowds may not produce reliable results (Hupont et al., 2014). However, Study 2 demonstrated through systematic replication that crowds can be supported to generate a labels with higher levels of agreement and contribute to a classifier with a higher level of accuracy. With the known limitation that the replication was subject to four differences detailed in section 5.2.2.

As I know SA is a context sensitive technology meaning that it is frequently more accurate when used in a context similar to the context where it was generated. This is a good reason to pursue further refinement of the approach of generating a classifier based on student opinions as it might enable more domain specific classifiers to examine

how the valence of text expressions relate to learning. Based on these results the approach may be the most suitable for massive open online courses where there are many students as they would produce more examples than I achieved in this study and that could facilitate examining the effects of scaling up the approach.

5.5 LIMITATIONS AND FUTURE WORK

One of the limitations of Study 2 is that I continue to use student examples as the benchmark for accuracy. While the results were encouraging these results continue to use the same data collection method as I used in Study 1. This is important in terms of replicating the results though it maintains a focus on aligning SA with student perceptions. While I might question the usefulness of student perception based on the questions raised around reliability it is important to remember that the general goal of SA is to detect the opinion of the author and reaction it elicits in the intended audience. Therefore, the limitation of questionable reliability of student opinion is a more general challenge that any SA method should consider when evaluating the accuracy of SA.

Study 1 established SSSAC Logistic as achieving the highest level of accuracy in predicting student labels from student sourced examples. Study 2 established that student sourcing SSSAC Logistic was improved in terms of inter-rater reliability of examples using emotional sentence starters and that combining data sets illustrated that this approach can reach higher levels of accuracy with more data. As Study 1 and Study 2 were limited by solely evaluating classifiers based on student sourced examples, in line with recommendations by Weidman et al. (2016) there is a need to examine how the measure correlates with other emotional measures. While these results are encouraging it is necessary to investigate how this measure might relate to other instruments. In Study 3 I shift the focus from cross-validation and replication as methods to validate SA to considering correlation with psychological instruments.

CHAPTER 6 STUDY 3 – EXPLORING THE EMOTIONAL JOURNEY OF STUDENTS FROM DISPOSITIONS, INCIDENTAL EMOTIONS, EMOTIONAL EXPRESSION, AND OVERALL EMOTIONAL EXPERIENCE



6.1 INTRODUCTION

In Study 3 I explore the validation of the measure SSSAC Logistic with a correlation approach which examines the extent to which SSSAC Logistic, a measure of emotion expression, correlates with established psychological measures of emotional experience. As Sentiment Analysis (SA) is used for a range of tasks it is important to understand exactly what this is capturing.

In order to examine validity, I consider the emotional experience from three dimensions, namely emotional expression, emotional state, and emotional trait. The first dimension is emotion expression. I consider how the SSSAC Logistic correlates with measures of emotional state and emotional trait.

The dimension of emotional state indicates the emotions students are experiencing at a given point in time. I use emotional state measures at the beginning and the end of the lab activity. The measure of emotional state at the beginning of the activity is intended to capture the emotional state going into the activity to measure how

emotions at the beginning might influence expression during the lab activity. Emotional states prior to an activity are thought to influence behavior and have been referred to as Incidental emotions which are emotions that may not be relevant to the current situation that still have influence over our judgements and behaviors (Lerner & Keltner, 2015). The second emotional state measure is at the end of the activity. The purpose for measuring emotional state at the end of the activity is to determine the extent to which emotional expression during the activity relate to the emotional state at the end of the activity. Effectively, if a student feels positive before and after the activity they might express more positive emotion during the activity. While emotional state before and after the activity may relate to emotional expression during the activity students may have emotional traits that also influence expression. One emotional trait related to emotional expression is the extent to which students are comfortable with emotion expression. Some students might be more likely to say positive things because of their personality. To unpack the emotional trait for emotional expression I first define a model of emotional expression. The consensual model of emotion suggests that after receiving some stimulus that is evaluated in a manner that generates an emotional experience there is a response modulation wherein dispositions for expression modulate emotional expression (Gross et al., 2000).

We previously reviewed studies which administered the Berkley Expressivity Questionnaire (BEQ) as a dispositional measure for emotion expression, which has the subcomponents of positive expressivity and negative expressivity (Gross et al., 2000; Kahn et al., 2016). We also reviewed studies which used PANAS (Drake et al., 2006; Gross et al., 2000; Kahn et al., 2016) which is commonly used in emotion research and it the instrument that best represent bivariate perspectives on valence (Green et al., 1993; Leue & Beauducel, 2011; Watson et al., 1988, 1999). PANAS produces largely independent scores for positive and negative affect (Feldman Barrett & Russell, 1998). A study which found a correlation between BEQ and PANAS (Gross & John, 1997a). We reviewed a studies which used Mixed Emotion Scale (MES) which aims at explicitly measuring the integrative nature of positive and negative valence and was shown to be distinguished as a measure from the related measures (Berrios & Totterdell, 2013) of ambivalence (Pekrun et al., 2011) and intolerance of ambiguity. A study which

used the emotional self-report tool React (Hillaire et al., 2016). We use these instruments to conduct correlation analysis within these measures and as compared with sentiment analysis to explore the extent to which psychological instruments related to SA. By using dispositional, incidental, and overall experience measures of emotion I first examine the extent to which there are correlations between measures of similar valence to establish if there are direct relationships between measures of the same category of valence (i.e., positive, negative, or mixed). To examine the extent to which there are inverse relationships between positive and negative measures, I explore evidence for integrative relationships, which are defined as a reciprocal relationship between positive and negative, where more positivity necessarily means less negativity, and vice versa. When considering previous correlational work on emotional measures weak correlations have been found. For example, considering how SSSAC Logistic in conjunction with established measures that have previously shown weak correlations this raises research question 5 (RQ5): To what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?

While research question 5 examines how SSSAC Logistic relates to a host of emotional measures I also conduct a similar analysis for the primary benchmark technology SentiStrength. The purpose for conducting this analysis is to determine the extent to which a general-purpose technology designed to measure sentiment in a similar fashion has correlation with psychological measures. As I have introduced the SSSAC it is necessary to examine how an alternative SA technology relates to psychometric instruments which raises research question 6 (RQ6): To what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?

In summary Study 3 examines the correlation between dispositions for emotion expression, incidental emotion, emotional expression, and the overall emotional experience with RQ5 and RQ6 by considering the extent to which there is correlates with SSSAC Logistic and SentiStrength. Both anticipated and unanticipated correlations are reported, where anticipated correlations are situations where measures of the same theoretical basis for valence have correlations. Unanticipated correlations are where

measures designed based on parallel theoretical models of valence correlate with measures based on integrative theoretical models of valence (or vice versa).

6.2 METHODS

6.2.1 SETTING

This Study examines the same setting as Study 1 and Study 2 focusing on the 2017 data collection comprised of both 2017SS and 2017C.

6.2.2 PROCEDURE

Study 3 examines input, process, and output emotional measures to examine correlation of the SSSAC Logistic measure validated in Study 2. Using the same 884 students who participated in a 60-minute lab in groups of 4 ($M=3.49$; $SD=0.89$) in 2017 as a part of Study 2, I focus now on correlation of the measurement SSSAC Logistic. A detailed breakdown of the Procedure is provided in Table 5.2.2. In this study I asked participants to answer the BEQ (Gross & John, 1997b) prior to participating in the study (see 6.2.4.1). The group task started with a 10-minute warm up activity where students participated in a statistics sampling activity. At the end of the warm-up activity students self-reported their emotional reaction using React (Hillaire et al., 2016). In the React interface students could select between 0 and 12 emotional words to describe their reaction to the activity. The interface has 6 positive words and 6 negative words (see also section 5.2.2.). During the group activity students participated in a group chat. The comments from the chat were classified as positive, negative, neutral, and mixed using the SSSAC classifier SSSAC Logistic (Hillaire et al., n.d.).

I administered positive affect negative affect schedule (PANAS) and the mixed emotion scale (MES) (Berrios & Totterdell, 2013) as exit surveys in Study 3. I used these exit surveys to determine the extent to which students self-reported the group learning experience as positive or negative (using PANAS), and the extent to which they described the experience as mixed emotion (MES). By using the PANAS to measure affect as bivariate, I incorporated as measure that anticipated parallel relationships

between positive and negative affect. By using MES I also incorporate a measure designed to consider the integrative nature of positive and negative affect. These two used in combination reflect the theoretical perspective that affect is both parallel and integrative (Cacioppo et al., 1999).

6.2.3 PARTICIPANTS

The participants in this study are the same as in Study 2 detailed in section 5.2.3.

6.2.4 INSTRUMENTS

6.2.4.1 BERKLEY EXPRESSIVITY QUESTIONNAIRE (BEQ)

In this study I asked participants to answer the BEQ (Gross & John, 1997a) prior to participating in the study. There are elements of the consensual model of emotion experiment (Gross et al., 2000) that are applicable to Study 3 as it demonstrates that predictions of expression from peers correlate with BEQ (expressivity) and PANAS (experience). In Study 3 I use a classifier based on student perception of emotion expression to predict student's emotion expression which is similar to having peers make predictions. There are also distinctions between consensual model of emotion study (Gross et al., 2000) and Study 3 in that the consensual model of emotion study used peers who knew the target participants for at least three years, it was conducted in North America, and it did not compare BEQ or PANAS scores with Sentiment Analysis.

I administered 9 items from the Expressivity questionnaire which were selected in consultation with the teacher of the respective course to identify the appropriateness of the question for the students. There were three items per construct: Comfort with expressing negative emotion (items: BEQ_NE3, BEQ_NE9, BEQ_NE13), comfort with expressing positive emotion (items: BEQ_PE10, BEQ_PE1, BEQ_PE6), and Impulsivity Strength (items: BEQ_IS11, BEQ_OS15, BEQ_IS12). The BEQ was administered and students provided responses between September 14th 2017 and October 4th 2017 with responses submitted between 4 weeks ahead of the lab to a few days ahead of the lab as the lab experiment was conducted between October 9th and October 13th 2017. The survey was sent to 1075 students and 963 students provided

responses. Of the 963 that provided responses in total 869 (90%) participated in the study. The loss of 94 responses is detailed in Figure 5.2.3 which outlines that 191 of the 1075 students did not participate in the study.

For the 855 responses, I first ran a confirmatory factor analysis which indicated that the model did not have an adequate fit as indicated by the RMSEA = 0.143 (90% CI [0.132, 0.155]) where a value below 0.08 is considered a good fit. The CFI was 0.843 indicated poor fit as it was below the cutoff of value of 0.90. As the model did not appear to be a good fit I next did an exploratory analysis

When comparing the three constructs with 9 items with a single construct model using the Omega function from the psych package in R I saw that BEQ_PE10 cross loaded on all three constructs. In addition to this cross-loading problem I found that two items loaded on the wrong factor. BEQ_NE3 loaded on factor 1 which is best described as Comfort with expressing positive emotion as it included BEQ_PE1 and BEQ_PE6. Similarly, BEQ_IS12 loaded on factor 3 which is best described as Comfort with expressing negative emotion as it included BEQ_NE9 and BEQ_NE13.

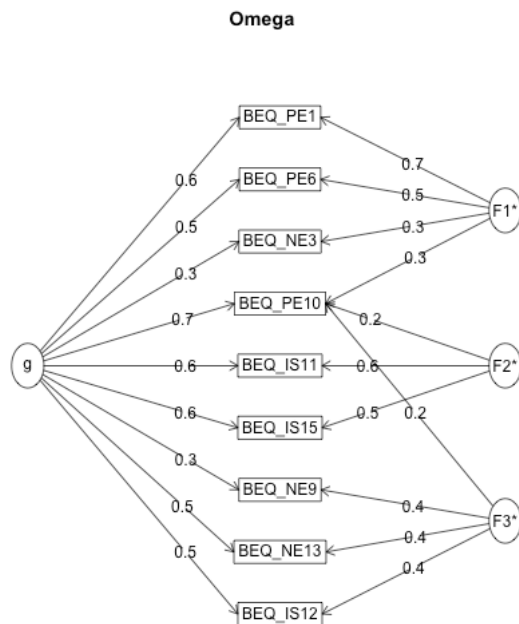


Figure 6.2.4.1a – BEQ three constructs compared to one construct for nine items

In terms of the items which loaded on the incorrect factor I examined the text of the item. BEQ_IS12 states: “I am sometimes unable to hide my feelings, even though I

would like to.” This text does not explicitly state positive or negative emotion and described hiding feelings which best aligns with Impulsivity Strength as a construct. BEQ_NE3 states: “People often do not know what I am feeling.” Which has previously loaded with the negative construct (Gross & John, 1997a).

Based on the cross-loading and loading on the incorrect construct I omitted three items: BEQ_PE10, BEQ_NE3, and BEQ_IS12 which left two items per construct. The remaining items by construct were as follows: Comfort with expressing positive emotion (items: BEQ_PE6, BEQ_PE1); Comfort with expressing negative emotion (BEQ_NE13, BEQ_NE9); Impulsivity Strength (items: BEQ_IS15, BEQ_IS11). Again, using the Omega function from the psych package in R I compared the three-factor model with two items per construct with the one factor model.

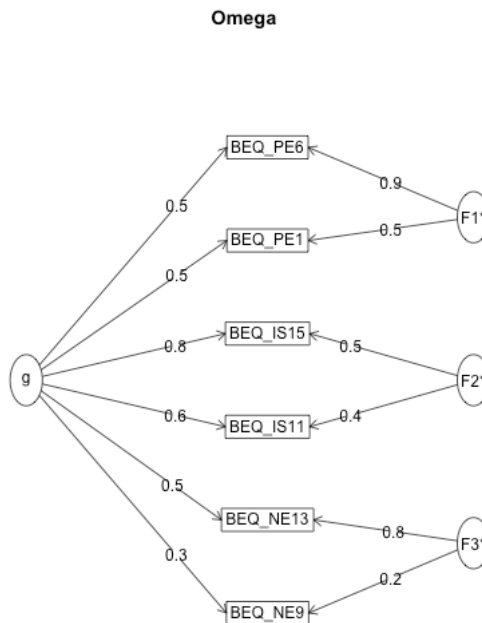


Figure 6.2.4.1b – BEQ three constructs compared to one construct for six items

The model using six items and three factors had a good fit as indicated by the RMSEA of 0.053 (90% CI [0.029, 0.079]) which placed the value below the cutoff of 0.08. In addition, the CFI was 0.99 which was above the cutoff of 0.90. All items were above the cutoff value of 0.3 except BEQ_NE9 which loaded with a value of 0.2 indicating that BEQ_NE9 was not a good fit with factor 3 (BEQ Negative).

Table 6.2.4.1a The Cronbach's alpha for the three constructs of the Berkley Expressivity Questionnaire

		α
BEQ	Positive	0.78
	Negative	0.50
	Impulsivity	0.79

The results of Cronbach's alpha indicated that BEQ Positive had a Cronbach's alpha of 0.78, BEQ Negative had a Cronbach's alpha of 0.50, and BEQ Impulsivity had a Cronbach's alpha of 0.79. Using the cutoff of 0.70 this indicated that BEQ Positive and BEQ Impulsivity items had reliable responses. However, BEQ Negative had a Cronbach's alpha of 0.50 indicating the responses between the two items were not reliable. While there are potential challenges for BEQ I examine dispositions for positive and negative expression and consider the challenges of the instrument as a factor when interpreting the results of correlation analysis.

Table 6.2.4.1b Berkley Expressivity Questionnaire items and text retained after exploratory factor analysis

Item	Text
BEQ_PE1	Whenever I feel positive emotions, people can easily see exactly what I am feeling.
BEQ_PE6	When I'm happy, my feelings show.
BEQ_NE9	No matter how nervous or upset I am, I tend to keep a calm exterior. (reverse)
BEQ_NE13	Whenever I feel negative emotions, people can easily see exactly what I am feeling.
BEQ_IS11	I have strong emotions.
BEQ_IS15	I experience my emotions very strongly.

As demonstrated in Table 6.2.4.1b the questions that remain for positive emotion have face validity as the items state: "Whenever I feel positive emotions, people can easily see exactly what I am feeling."; "When I'm happy, my feelings show." Similarly, the items for negative expression have face validity as the items state: "No matter how nervous or upset I am, I tend to keep a calm exterior." (reverse); "Whenever I feel negative emotions, people can easily see exactly what I am feeling". However, when looking at the items which remain for Impulsivity Strength the remaining items do not

mention regulation: “I have strong emotions.”; “I experience my emotions very strongly”. The item for impulsivity strength which was omitted stated: “I am sometimes unable to hide my feelings, even though I would like to.” Given that the item exclusion appears to have characteristically changed the Impulsivity Strength construct this construct was omitted from further analysis. While the negative construct had a low level of reliability (see Table 6.2.4.1a - Cronbach’s alpha 0.50), the items had face validity and contributed to an overall model with good fit statistics (see Figure 6.2.4.1b).

Finally, I formally tested the normality of the BEQ scores for positive and negative using the Shapiro-Wilk’s test. Both had significant differences from the normal distribution. To examine the distributions a histogram for the scores of each constructed are presented. The histograms display that the majority of students had a disposition for positive expression as depicted in figure 6.2.4.1c there is a left skewed distribution of BEQ Positive Scores. Students also had an aversion towards negative expression as depicted in figure 6.2.4.1d there is a right skewed distribution of BEQ Negative Scores.

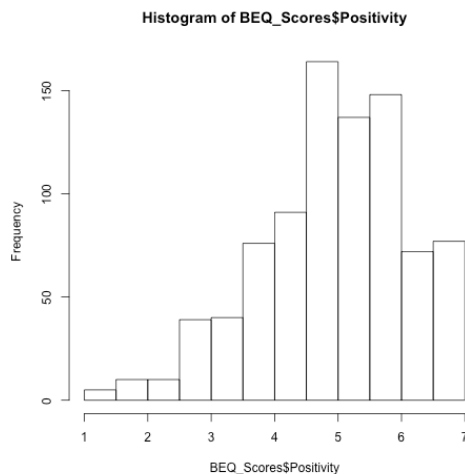


Figure 6.2.4.1c – Histogram of BEQ Positive scores

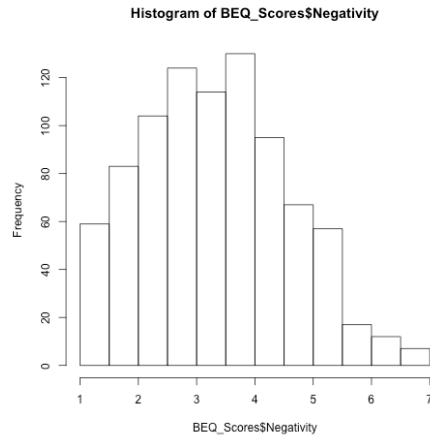


Figure 6.2.4.1d – Histogram of BEQ Negative scores

In summary, 855 participants in the study were administered 9 items from the BEQ for three constructs and a confirmatory factor analysis revealed a poor fit (RMSEA = 0.143 (90% CI [0.132, 0.155]); CFI = 0.843). I followed up with exploratory factor analysis which confirmed that 3 factors fit the six items (RMSEA = 0.053 (90% CI [0.029, 0.079]); CFI = 0.99). When examining how the items loaded onto the constructs I eliminated 3 items which left 2 items for each of the three constructs (see remaining items in Table 6.2.4.1b). There were reliable responses for the constructs of Positive ($\alpha=0.78$) and Impulsivity ($\alpha=0.79$). While the responses for the Negative construct was not reliable ($\alpha=0.50$) likely due to the item BEQ_NE9 as the loading was 0.2 below the cutoff value of 0.3. While both Positive and Negative constructs of the BEQ kept face validity the resulting scores for both constructs had a significant difference from the normal distribution (see Figure 6.2.4.1c and Figure 6.2.4.1d). Based on the fit of the model for three constructs and six items and the review of the face validity of the items I determined that BEQ Positive and BEQ Negative constructs were suitable for further analysis acknowledging that BEQ Negative had some challenges with reliability. The BEQ was then used in correlation analysis (See section 6.3) to see if participants more likely to express positive emotion and/or negative emotion more frequently had messages coded as positive or negative by sentiment analysis.

6.2.4.2 REACT

As previously indicated in Study 2, students participated in a group assignment by first doing a warm up activity for 10 minutes. The activity was based on the mathematics and statistics course concepts and was designed using materials provided by the instructor of the course. The content of the activity was time constrained to 10 minutes creating a demanding warm up exercise. Participants were told at the end of the 10 minutes to participate in a group exercise. The group exercise first models how to use the React feature to report your reactions to the warm up exercise. Then students self-reported their emotional reaction using React (Hillaire et al., 2016). In the React interface students could select between 0 and 12 emotional words to describe their reaction to the activity. The interface has 6 positive words and 6 negative words (see Figure 5.2.2a - Interface of 'React' to self-report emotional response). These words have been previously been interpreted in terms of positive and negative valence demonstrating a correlation with learning (Hillaire et al., 2018).

I used the same scoring method of substituting valence scores for each term based on the Warriner et al. (2013) SA dictionary as used in previous research with React (Hillaire et al., 2018). To see the Valence scores substituted for React responses refer to table 6.2.4.2.

Table 6.2.4.2 React Word Valence Scores

React Option	Warriner Dictionary	Valence Score
CURIOUS	curious	6.58
GOOD	good	7.89
INTERESTING	interesting	6.78
ENGAGING	engaged	6.78
CALMING	calm	6.89

BORING	boring	2.71
SAD	sad	2.1
ANNOYING	annoying	3
CHALLENGING	challenge	5.95
CONFUSING	confusion	3.32
FRUSTRATING	frustrating	2.57
DULL	dull	3.4

By substituting valence scores for words selected and averaging the scores the results were a continuous measure with a non-normal distribution. The distribution generated had a Min = 2.1; Max =7.89, Mean = 5.32, Median = 5.3, SD = 1.27. The distribution of scores are illustrated in Figure 6.2.4.2.

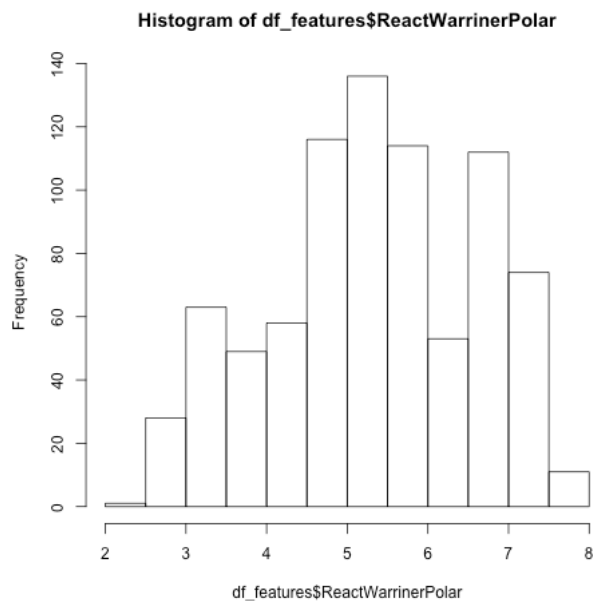


Figure 6.2.4.2 – Histogram of React Scores Reporting Incidental Emotions to the Warm-up Activity

6.2.4.3 GAG-OF-WORDS ENSEMBLE VALENCE (SSSAC LOGISTIC)

I predicted the valence category of positive, negative, neutral and mixed for each communication using SSSAC Logistic. For each student I calculated the percentages of communication for each category. The accuracy for those predictions were reported in Study 2.

6.2.4.4 MIXED EMOTION SCALE (MES)

During the exit survey students answered the Mixed Emotional Scale (MES), a four-item questionnaire designed to detect the extent to which a person has experienced mixed emotion. In this study I asked participants to answer the Mixed Emotion Scale (MES) (Berrios & Totterdell, 2013) after participating in the study. I administered 4 items from the questionnaire which were selected in consultation with the teacher to identify the appropriateness of the question for the students. There were four items for the one constructed of mixed emotion (ME1, ME7, ME10, ME12). The survey was sent to 886 students who participated in the study (see Figure 5.2.3) and 756 (85%) students provided responses. The loss of 130 responses were a result of students who participated in the study that did not complete the exit survey items for the MES.

For the 756 responses, I first ran a confirmatory factor analysis which indicated that the model had an adequate fit as indicated by the RMSEA = 0.055 (90% CI [0.012, 0.104]) where a value below 0.08 is considered a good fit. The CFI was 0.998 which was also above the cutoff of value of 0.90 which indicated fit. Given that I had adequate fit I next examined how the factors loaded on the construct by using Omega to compare the one-factor model with a two-factor model. The results indicated that all items had a score above 0.3 (see figure 6.2.4.4a)

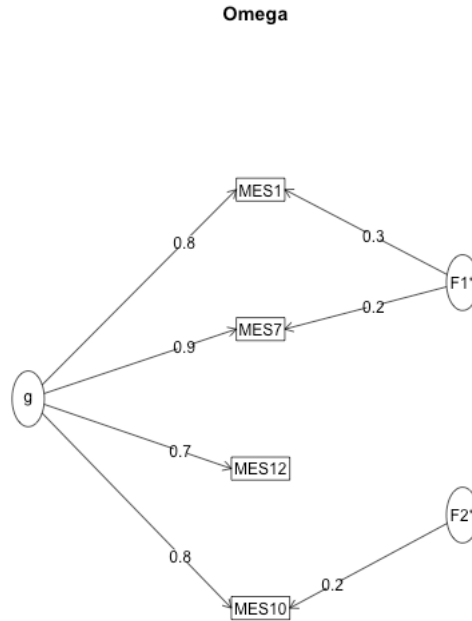


Figure 6.2.4.4a Comparing a one-factor model to a two-factor model of the Mixed Emotion Scale

Comparing with previous findings the factors had higher loading values. MES1 previously loaded with 0.78; MES7 previously loaded with 0.83; MES10 previously loaded with 0.88; and MES12 previously loaded with 0.61 (Berrios & Totterdell, 2013). In this study MES1 loaded with 0.8; MES7 loaded with 0.9; MES10 loaded with 0.8; and MES12 loaded with 0.7. The four items were reliable with a Cronbach's Alpha = 0.90 as compared with previously reported Cronbach's Alpha = 0.85 (Berrios & Totterdell, 2013). Table 6.2.4.4a provides the text for the items.

Table 6.2.4.4a Mixed Emotion Scale (MES) items and text retained after exploratory factor analysis

Item	Text
ME1	I felt a mixture of emotions.
ME7	I felt different emotions occur very quickly one after another.
ME10	I felt only one thing throughout the event or experience. (reverse)
ME12	I felt either positive or negative emotions but not both at the same time. (reverse)

Finally, I formally tested the scores for normality and found the MES scores were significantly different from normal as depicted in the histogram for MES scores (see figure 6.2.4.4b).

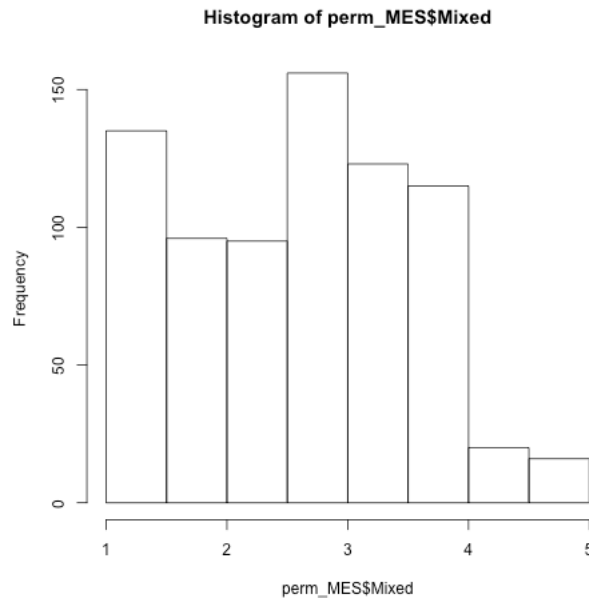


Figure 6.2.4.4b – Histogram of MES Mixed Scores

6.2.4.5 PANAS

In addition to MES, participants the 12-item PANAS questionnaire designed to determine the extent to which students experienced positive affect and negative affect. During the exit survey. Students answered the PANAS after participating in the study. I administered 12 items from the questionnaire where participants rated 12 emotion words on a scale from 1 to 7 to describe the overall lab experiment. There were six items for each of the two constructs Negative (Guilt, Fear, Sadness, Hostility, Shyness, Fatigue) and Positive (Serenity, Attentiveness, Self-Assurance, Joviality, Surprise). The survey was sent to 886 students who participated in the study (see Figure 5.2.3) and 756 students provided responses. The loss of 130 responses are the result of students who participated in the study that did not complete the exit survey items for the PANAS.

For the 756 responses, I first ran a confirmatory factor analysis which indicated that the model had an adequate fit as indicated by the RMSEA = 0.072 (90% CI [0.063, 0.082]) where a value below 0.08 is considered a good fit. The CFI was 0.914 which was also above the cut-off value of 0.90 which indicated fit. Given that I had adequate fit I next examined how the factors loaded on the construct by using Omega from the R

package ‘psych’ to compare the one-factor model with a two-factor model. The results indicated that all items had a score above 0.3 for their construct (see figure 6.2.4.5a). The item ‘surprise’ loaded on both positive and negative factors.



Figure 6.2.4.5a – Item loading for 12 item PANAS scale comparing one-construct to two-constructs

I removed the item ‘Surprise’ as it cross loaded which slightly improved the model fit resulting in the RMSEA = 0.064 (90% CI [0.054, 0.076]) where a value below 0.08 is considered a good fit. The CFI was 0.941 which was also above the cut-off value of 0.90 which indicated fit. The model without surprise is displayed in Figure 6.2.4.5b

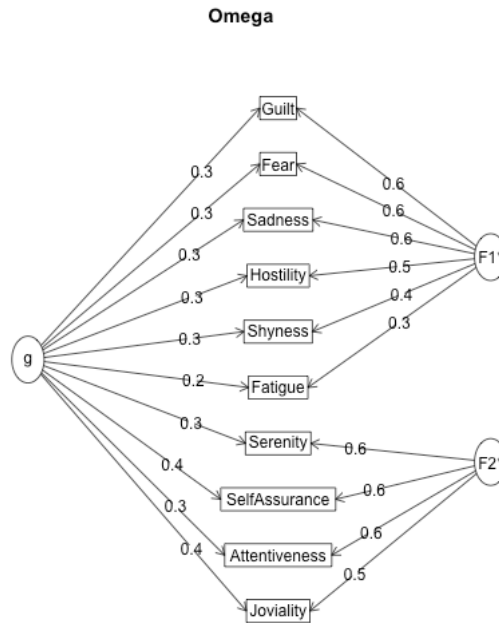


Figure 6.2.4.5b – Item loading for 11 items PANAS scale comparing one-construct to two-constructs

Finally, I checked both constructs for normality and found that both Positive and Negative constructs were significantly different from the normal distribution as demonstrated by the histograms (see Figure 6.2.4.5c and Figure 6.2.4.5d)

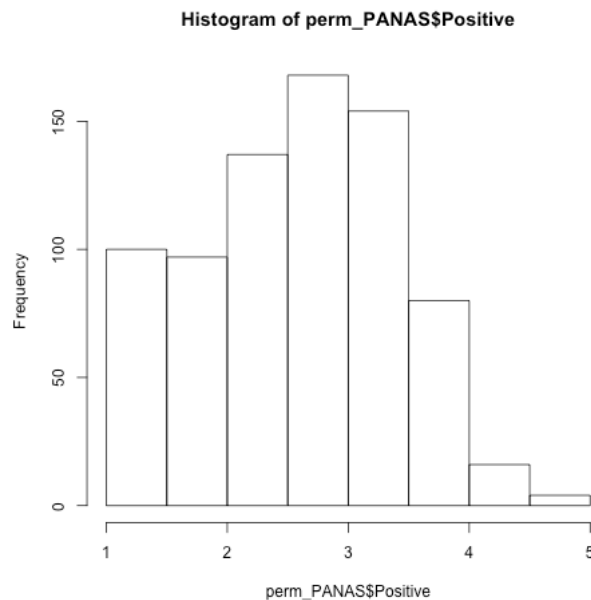


Figure 6.2.4.5c – Histogram of PANAS Positive Scores

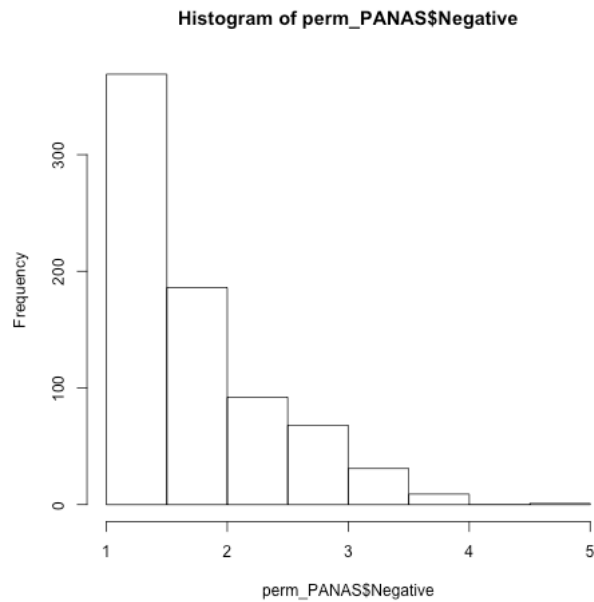


Figure 6.2.4.5d – Histogram of PANAS Negative Scores

6.3 ANALYSIS

I used a correlation matrix to examine BEQ, React, SSSAC Logistic, PANAS, and MES with SSSAC Logistic(2106). I used a Pearson correlation and reported the 95% confidence intervals for the correlations. The rationale for the correlation matrix was to identify convergent and discriminate validity for SSSAC Logistic by examining the correlations with the instruments BEQ, React, PANAS, and MES. I then conduct a similar correlation analysis between SentiStrength with the same measures to see the extent to which there is correlation between a comparable SA technology with the same instruments (BEQ, React, SSSAC Logistic, PANAS, and MES).

6.4 RESULTS

6.4.1 RQ5: TO WHAT EXTENT ARE THERE CORRELATIONS BETWEEN EMOTIONAL EXPRESSION MEASURED BY A STUDENT SOURCED SENTIMENT ANALYSIS CLASSIFIER, STATES OF EMOTION, AND TRAITS OF EMOTION?

The expectations for correlation outlined in the introduction indicated that a correlation coefficient of 0.25 would align with previous correlation of Sentiment Analysis technologies with psychological measures. As indicated in our results (see Table 6.4.1) the only measure which had a significant relationship with SSSAC Logistic was the BEQ sub-construct of disposition for positive expression with an inverse relationship to expressions of mixed emotion ($r = -0.09$, $p < 0.01$). These results indicate a lack of correlation with emotional states measured with React, PANAS, and MES. As the only correlation found was not near the expectation 0.25 these results indicate there is a lack of correlation between expression measured by SSSAC Logistic and the measures of emotional state and trait. Furthermore, the trouble I had with factor loadings of questions from the BEQ make the weak correlation suspect and insufficient to be considered as correlation on its own.

In addition to the trouble with BEQ I can bring into question if SSSAC Logistic had a sufficient level of accuracy to adequately examine questions of correlation., As I reported in Figure 4.3.3 SSSAC Logistic had a recall for positive of 0.52, negative of 0.48 and mixed of 0.30. To consider the recall rate as a factor in the analysis this puts further doubt that the correlation between the BEQ and SSSAC Logistic is evidence of correlation as the correlation with SSSAC Logistic was with mixed expression which had the lowest recall (i.e., 0.30). There is a reasonable chance that the correlation found is related to measurement error for SSSAC Logistic, BEQ, or both instruments.

At the same time, it is possible that if the recall for mixed expression improved or if the reliability of the BEQ was better, that the correlation might be stronger and provide some evidence of correlation., One distinct difference between mixed emotion and positive and negative emotions is that the co-activation of positive and negative is anticipated unstable and shorted lived (Larsen, McGraw, & Cacioppo, 2001). While

mixed emotions may be a briefer experience as compared with positive and negative emotion, this brief existence may best be measured in shorter intervals such as the context of this study (a one-hour lab activity). Conversely stated the length of the activity of this study may be too short to adequately model how positive and negative emotion expression relate to experience. For example, SA that predicted a single scale from positive to negative expression was found to have a relationship with learning outcomes when examining data collected over a 7-month period (Hillaire et al., 2018).

The conclusion that I can draw from this evidence is that in terms of answering research question 5 there is no evidence to suggest correlation between SSSAC Logistic and the psychological instruments administered during the one-hour lab activity. Future work should consider how the definition of SA as the opinion of the author and reactions it elicits from the intended audience may require measures that consider social interaction as potentially better instruments to examine correlation., However, given that SA should in part reflect the opinion of the author, which might relate to a personal emotional experience, the evidence in Study 3 suggests that this relationship may not exist or if it does then it may at least be obfuscated because the opinion of the author is blended with the reactions of the intended audience. Future work might consider training a SA algorithm solely on self-reflection of authors on their own post to determine correlates with psychological measures of emotion. This would of course be in conflict with the expressed goal of SA but it would provide further insight into the challenges associated with SA and correlation.,

The first limitation to this study is that the measures selected for correlation analysis had issues in terms of validation as the PANAS and BEQ did not confirm the expected factor loading. BEQ had particularly difficult challenges. The difficulty I had with BEQ is reflected by other studies which had difficulties of a similar nature.

Another limitation to this work is that I administered React ahead of the group work activity and PANAS after the group work activity. It would have been ideal to use the same instrument as a pre-test and post-test for validation. React had a positive correlation with PANAS for the positive score and an inverse correlation with PANAS for the negative score. These relationships are expected as react produced a single score where lower values indicate a negative reaction while higher values indicate a positive

reaction. It would be difficult to use React and PANAS to examine questions about transitions between emotional states given that they had weak correlations in terms of correlation., While they are different instruments the convergence between them demonstrated that there is some validity to consider them as related measures. However, the use for this study is to consider how SSSAC Logistic correlated with either of the measures as evidence of correlation.,

Table 6.4.1 Correlations with confidence intervals for SSSAC Logistic and comparison measures

Variable	1	2	3	4	5	6	7	8
1. BEQ Positive								
2. BEQ Negative	.31** [.25, .38]							
3. React	-.08* [-.15, -.00]	-.03 [-.10, .04]						
4. SSSAC Logistic Positive	.07 [-.00, .14]	.04 [-.03, .11]	-.00 [-.07, .07]					
5. SSSAC Logistic Negative	.01 [-.07, .08]	.05 [-.02, .12]	-.03 [-.10, .04]	-.27** [-.34, -.20]				
6. SSSAC Logistic Mixed	-.09* [-.16, -.01]	.01 [-.06, .09]	.05 [-.02, .12]	-.15** [-.22, -.08]	-.24** [-.31, -.17]			
7. PANAS Positive	.02 [-.05, .10]	.03 [-.04, .10]	.10** [.03, .17]	-.01 [-.09, .06]	-.07 [-.14, .01]	.04 [-.03, .12]		
8. PANAS Negative	.05 [-.02, .13]	.09* [.01, .16]	-.14** [-.21, -.07]	.02 [-.05, .10]	-.01 [-.08, .06]	-.03 [-.11, .04]	.23** [.16, .29]	
9. MES Mixed	.08* [.00, .15]	.10** [.03, .17]	-.04 [-.11, .03]	.04 [-.03, .11]	-.05 [-.12, .02]	.07 [-.00, .14]	.25** [.18, .32]	.33** [.26, .39]

Table 6.4.1 continued

Variable	1	2	3	4	5	6	7	8	9
<i>M</i>	5.11	3.54	5.35	0.25	0.20	0.08	2.68	1.7 6	2.72
<i>SD</i>	1.21	1.27	1.26	0.17	0.17	0.11	0.86	0.6 8	1.01

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

6.4.2 RQ6: TO WHAT EXTENT ARE THERE CORRELATIONS BETWEEN EMOTIONAL EXPRESSION MEASURED BY SENTISTRENGTH, STATES OF EMOTION, AND TRAITS OF EMOTION?

As indicated in our results (see Table 6.4.2) the only measure which had a significant relationship with SentiStrength was that React which had a positive correlation with expressions of mixed emotion ($r = 0.10$, $p < 0.01$). These results indicate a lack of correlation with emotional state measures React, PANAS and MES. As the only correlation found (BEQ and SentiStrength) was not near the expectation 0.25 these results indicate there is a lack of correlation between expression measured by SentiStrength and the measures of emotional state and trait.

It is important to first recall that SentiStrength was predicted to have a lower overall accuracy than SSSAC Logistic in the benchmarks reported in Study 1 (see Figure 4.3.3) as I anticipated the recall for SentiStrength to be a positive recall of 0.54, a negative recall of 0.33, and mixed recall of 0.15. Given that mixed is the only correlation found in the correlation analysis it is important to underscore that SentiStrength was predicted to have a low recall of 0.15. This may indicate that the correlation I found is explained by measurement error. The best interpretation of these results is to conclude that SentiStrength did not have evidence of correlation.,

It is important to note that SentiStrength distinguishes the measurement goal away from emotional measurement by stating that the target of measurement for SentiStrength is a measure that can reflect the authors internal state, the intended message interpretation, or the reader's internal state (Thelwall et al., 2010). While this aligns with the definition of SA as a measure that captures the opinion of the author and the reactions it elicits in the reader, it supports the evidence that SA is perhaps not the best measure to use when seeking a measure that provides insights into the emotion of students. However, there is not clear direction on how to exactly validate a measure intended to capture a blend of the internal state of the author and the internal state of the intended audience. Again, I again reiterate the comparing convergence with social measures is out of scope for this thesis as the emphasis is on examining the extent to which there is correlates with psychological measures of internal emotional state and trait.

Table 6.4.2 Correlations with confidence intervals for SentiStrength and comparison measures

Variable	1	2	3	4	5	6	7	8
1. BEQ Positive								
2. BEQ Negative	.31** [.25, .38]							
3. React	-.08* [-.15, - .00]	-.03 [-.10, .04]						
4. SentiStrength Positive	.05 [-.02, .13]	.03 [-.04, .11]	.02 [-.06, .09]					
5. SentiStrength Negative	.01 [-.07, .08]	-.00 [-.08, .07]	-.03 [-.11, .04]	-.14** [-.21, - .07]				
6. SentiStrength Mixed	-.05 [-.12, .03]	-.02 [-.10, .05]	.10** [.02, .17]	-.16** [-.23, - .09]	-.04 [-.11, .04]			
7. PANAS Positive	.02 [-.05, .10]	.03 [-.04, .10]	.10** [.03, .17]	-.02 [-.10, .05]	-.03 [-.11, .04]	.00 [-.08, .07]		
8. PANAS Negative	.05 [-.02, .13]	.09* [.01, .16]	-.14** [-.21, - .07]	-.04 [-.12, .03]	-.03 [-.07, .08]	-.02 [-.10, .05]	.23** [.16, .29]	
9. MES Mixed	.08* [.00, .15]	.10** [.03, .17]	-.04 [-.11, .03]	-.05 [-.12, .03]	.01 [-.07, .08]	-.01 [-.09, .06]	.25** [.18, .32]	.33** [.26, .39]

Table 6.4.2 continued

Variable	1	2	3	4	5	6	7	8	9
<i>M</i>	5.11	3.54	5.35	0.25	0.20	0.08	2.68	1.76	2.72
<i>SD</i>	1.21	1.27	1.26	0.17	0.17	0.11	0.86	0.68	1.01

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

6.5 CONCLUSIONS

To answer to RQ5: “To what extent are there correlations between emotional expression measured by a student sourced sentiment analysis classifier, states of emotion, and traits of emotion?”, there is nearly no evidence of correlation between SSSAC Logistic and any of the measures for states and traits of emotion. There was one weak inverse correlation found between BEQ Positive and SSSAC Logistic Mixed ($r = -.09$, $p < 0.05$) indicating that the more comfortable a student was with expressing positive emotion the less likely they were to express mixed emotion. Given that the correlation was below the expectation of 0.25 I conclude there is insufficient evidence for correlation between emotional expression, emotional states, and emotional states.

To answer to RQ6: “To what extent are there correlations between emotional expression measured by SentiStrength, states of emotion, and traits of emotion?”, there is nearly no evidence of correlation between SentiStrength and any of the measures for states and traits of emotion. There was one weak correlation found between React and SentiStrength Mixed ($r = .10$, $p < 0.01$) indicating that the more when students report a positive emotional state after the warm up activity they are more likely to express mixed emotion. Given that the correlation was below the expectation of 0.25 I conclude there is insufficient evidence for correlation between emotional expression, emotional states, and emotional traits.

As neither of the SA technologies developed using the ESM demonstrated correlation with any of the measures used (BEQ, React, PANAS, MES) I conclude that

emotion expression detected using SA developed based on ESM does not provide insight into the emotional experience of students. One potential explanation for this result is that SA is not accurate enough to measure convergence with psychological measures of emotion. However, the weak correlations found were with the emotional category of mixed which I anticipate has the lowest recall (see Figure 4.3.3 which predicts SSSAC Logistic to have recall of 0.30 for mixed and SentiStrength to have a recall of 0.15). However, SSSAC Logistic had a predicted recall 0.52 for positive expression and SentiStrength had a predicted recall of 0.54 for positive expression. If accuracy were the main cause then the entire report of recall across all twelve benchmarks reported in Figure 4.3.3 would be rather condemning for the use of the majority of SA technologies as the majority of predicted accuracy for the benchmarks was at or below 0.52. When taking an accuracy interpretation on these results the suggestion would be to avoid using SA in the context of learning until there is a validation of any SA measure in the context of learning.

An alternative to the accuracy argument is that the lack of convergence between measures is indicating that emotional expression is characteristically different from emotional experience. In the social theory of emotion proposed by Barrett she suggests that emotional expression is used to express emotional experience or as a means of social influence. It is possible that students are using emotional expression for both purposes and in doing so obscure any potential relationship with psychological measures of emotional state or trait. It is in fact even possible that in the context of a one-hour group lab activity the majority of emotional expression is related to social regulation. It is not in the scope of Study 3 to examine this facet of the theoretical model of social emotion so the extent to which communication focused on social regulation not examined. In fact, it is possible that some alternative model of emotion is more appropriate for this context. The theory of Situated Affectivity which focuses on how emotions relate to goal states may be more appropriate for learning.

As SSSAC Logistic is a measure built based on the students interpreting comments from their own group discussions and provides predictions that do a better job of matching student perceptions than any of the benchmarks used in Study 1 I may also conclude that students do not have a good line of sight on the emotional states of

their peers when making predictions based on text communication from their peers. Effectively, emotion expression in text appears to be inadequate to understand the emotional state of the author. If this is the case then there is a potential criticism on the sampling method used to train the classifier. It may have been more effective to gather sample communications labelled by the author and consider that to be ground truth rather than getting labelled data that reflects opinions of the author and the intended audience. However, that sampling technique closely aligns with a common definition of SA which considers the measure to be a combination of the intent of the author and the reactions communication elicits from the intended audience. So, it is possible that SA is a measure that speaks more to the social context than it does to the emotional state of any individual in the context.

In summary SA as examined in Study 3 using a student sourced classifier and a state-of-the-art technology (both based on the ESM) are not effective measures for the emotional state of the authors of text-based communication. Further work should be done to increase the accuracy of the measure, which might include departing from the common definition of SA, to generate a measure capable of providing insight into the emotional state of students. In addition, these results indicate that it may be necessary to explore psychological measures focused on social aspects rather than emotional state may be necessary to gain clear insight into exactly what SA is measuring in terms of emotion. It is also possible that in this context emotional expression and emotional experience are simply two distinct constructs.

6.6 LIMITATIONS AND FUTURE WORK

There were challenges associated with confirming the instruments used for correlation analysis - most notably with the BEQ. One of the limitations of this study is that the results are interpreted in conjunction with the potential problems with the instruments used in Study 3. In addition to challenges with the instruments used for correlation analysis the SA tools that I examine also have some issues in terms of their accuracy as reported in Study 1 (see Figure 4.3.3). As the results of RQ5 and RQ6 show no compelling evidence for correlation between emotional expression, emotional states, and emotional traits the safest interpretation is that the results suggest no evidence of

correlation with psychological emotional measures and the SA technology evaluated (SSSAC Logistic and SentiStrength).

I had issues with confirmatory factor analysis in this study it is possible the lack of correlations is explainable as measurement error in the instruments I used to test correlation in Study 3. It is possible that if I used surveys which had been confirmed when confirmatory factor analysis was run that more correlations would have been found. While there were challenges that were the most pronounced with BEQ, the PANAS instrument was close to the original set of factors. While challenges were found when interpreting the instruments used for correlation analysis this does not seem like the most likely explanation for the results of Study 3.

In future work, the most productive direction to consider an application for SA in the context of learning might be to investigate the social aspects of emotion. As the CTE outlines that communication can be to express emotion or as a social regulation strategy a logical next step would be to investigate the extent to which SA correlates with measures of social regulation of emotion. It may be useful to pair predictions of valence in text with a measure that determines the subject of the communication. For example, it might be necessary to distinguish between communication that is intended to indicate emotions to peers from communication that is intended to regulate peers given that both forms of communication are anticipated based on the CTE. Future work should also explore the extent to which SA has a relationship with students' goal states as SA may be a measure that is better explained by SAT rather than the CTE.

CHAPTER 7 CONCLUSION

7.1 INTRODUCTION

This chapter summarizes key findings from Study 1, Study 2, and Study 3. I then articulate the unique contribution of this work, outline limitations, and suggestions for future work. Finally, I provide guidance for researchers considering the use of sentiment analysis in the context of learning.

7.2 FINDINGS

In Study 1, I set out to create a new sentiment analysis (SA) measure by training a classifier using positive, negative, neutral, and mixed examples of expression provided by 767 students. The results in terms of accuracy demonstrated this approach was more effective than using Mechanical Turk labels or general approaches providing evidence that the method of student sourcing examples is a viable approach toward training a SA classifier. This provides more exciting and interesting insights of how to align SA with the opinions of students. This also raised some concerns about using existing SA technology in the context of learning as student agreement was moderate and if the goal of SA is to measure student opinion the level of student agreement presents a challenge for all SA work including this thesis.

In Study 2, I found that emotional sentence starters, a novel emotion aware tool proposed and evaluated in this study, supported students to achieve higher inter-rater reliability when providing examples and that more data and higher quality data improve the accuracy of the classifier. This suggests the results of this thesis may not have found the upper bound of potential accuracy. Furthermore, by increasing the accuracy the findings demonstrate substantial promise for supporting emotional communication during group work. This has implications beyond the scope of generating classifier as there may be unintended benefits to raising emotion awareness in students which were not explored in this thesis.

While Study 1 and Study 2 demonstrated potential for a SSSAC the scope of those studies was examining the accuracy of the SSSAC indicating that this approach

resulted in better accuracy than using mechanical turk ratings and general approaches. The core question about the measure is how it might apply to research in the educational context. To examine the potential it might relate to other psychological measure I conducted Study 3.

In Study 3, I set out to examine the extent to which a SSSAC demonstrated correlation with measures of state and trait emotion and emotion expressivity (PANAS, MES, React, and BEQ) to determine if emotion expression detected by the classifier demonstrated correlations. The results of the correlation analysis between the SSSAC and the emotional measures indicated no relationship between the valence of text expression of students with measures of their emotional experience. I confirmed the lack of evidence for emotional measures and SA by also examining correlation for a general SA classifier SentiStrength. SentiStrength also demonstrated a lack of correlation with other emotional measures.

While Study 1 and Study 2 provide reason to consider the SSSAC of higher accuracy, Study 3 introduced very interesting results which merit further expansion. Given that there appears to be a divide between SA measures and commonly used self-report measurements of state and trait emotion (both the student sourced measure introduced in this PhD thesis and the general technology), it raises questions about exactly how SA measures relate to emotion. Given that these results indicate that emotion expression as perceived by students from the context does not correlate with psychometric measures of emotion I must return to the theoretical basis for emotion expression to understand the divide.

Using the Component Process Model (CPM) of emotion (Scherer, 2009) I took the position that many components to emotion are associated with the unconscious (see Figure 2.1.1), as physiological symptoms, appraisal process, motor expression, and action tendencies are related to the unconscious reflection and regulation of emotion. From the perspective of CPM valid self-report of emotion occurs where conscious representation and regulation intersects with unconscious representation and regulation, and both intersect with verbalization and expression of emotion. I investigated this potential intersection of valid self-report of emotion using psychometric measures (PANAS, MES, React, and BEQ) and SA measures, and found a lack of correlation

between self-report of emotion and SA (see Table 6.4.1 and Table 6.4.2). From CPM I acknowledged that there are forms of emotional verbalization and expression that are likely not to represent the internal state of emotion. For example, where verbalization and communication intersect conscious reflection and regulation, but does not intersect with unconscious representation and regulation, I labelled this as regulated expression (see Figure 2.1.1). Using this theoretical frame to interpret our results it suggests that there may be a sufficient amount of regulated expression to make the detection of relationships with validated self-report measures challenging (as evidenced by our lack of correlations between a variety of measures and SA).

Another potential explanation of this finding could be related to social coordination and goal orientation. If social coordination between members of an online group is the primary reason for regulated expression, then what I might be observing in these findings is that students are focused on communication that emphasizes the coordination with peers over the expression of internal state. Barret articulates that emotions occur at collective intentionality (Feldman Barrett, 2017, pg. 139), where people in a social context like an online chat have to build a consensus about how to interpret emotional communication. It is possible that in verbalization and communication of emotion these students are orchestrating, but have not yet arrived, at a collective intentionality. For example, in Table 4.3.4a Student_05 identified the following message as negative: “i think [Country_05] has the highest percentage for children out of school”, while both SSSAC Logistic and Sentistrength SA technologies considered it to be neutral. This might be an example of a student seeking a collective intentionality to view Country_05’s education system in a negative way. As the goal of the group work exercise was to make a funding decision, working towards collective intentionality that considers Country_05’s education system could also be considered goal oriented in terms of the group assignment.

While social coordination goals may be described as working towards collective intentionality, these students may be focused on goals that are completely independent of collective intentionality. Our literature review of theoretical models of emotion included Situated Affectivity (Wilutzky, 2015) which suggests that to interpret emotion expression I should examine the goal of students. Through understanding the goal, I can

make an interpretation of their expression. From the SAT of emotion this intersection between goal and expression is what defines emotion. Given that there is a strong theoretical basis to consider the intersection between goal orientation and emotional expression and verbalization, it merits some expansion in the interpretation of these findings. If goal orientation explains the primary reason for regulated expression, then what I might be observing in these findings is that students use emotional expression as a means of achieving a goal associated with the group assignment, rather than communicating their internal emotional experience. For example, in Table 4.3.4a two communications which were detected as positive by students and SA technologies is “yeah [Student_03] your right! So I have our answer..” and “Yes so for now is [Country_04] !!!”. These comments demonstrate that both SA and students considered progress in the assignment as positive. If the goal of students is to complete the assignment, then these example communications might be ideal examples of positive emotion expression in this group assignment context. While it appears to have a clear relationship with the assignment goal, it is important to recall that I did not find a correlation between positive expression and any of the validated measures. This lack of correlation may suggest that positive expressions associated with goal achievement may be disjoint with internal emotional experiences.

It is of course possible that both social coordination and goal orientation are factors in explaining the division between self-report of emotion and SA measures. In fact, as our correlation analysis demonstrated a lack of evidence across all measures it might be reasonable to consider that multiple factors contribute to this result. Beyond the factors described in terms of regulated communication, another possible factor is that what is detected as emotional communication by SA is in fact disconnected communication. Previously I defined disconnected communication as verbalization and communication of emotion that does not intersect unconscious representation and regulation or intersect with conscious reflection and regulation. Barret provided an example of this (Feldman Barrett, 2017, pg. 139) as someone stomping their feet while walking to get dirt off of their shoes. Someone observing this may interpret this communication (arguably incorrectly) as negative emotion expression. Going back to Table 4.3.4a I can consider the communication “hi my name is [Student_01] I am from

[Country_04] and I like pizza”. In this message a student identified the communication as neutral and two SA technologies detected the message to be positive. While this communication may sound positive, this respective chat message may have little to do with the internal state of the author of the message. If disconnected communication helps to explain the divide between self-report and emotional verbalization and communication detected using SA, then text messages detected as emotional communication represents the mis-interpretation of the measure. Students may simply use positive and negative terms to describe something for which they have no associated internal emotional experience.

As part of the emphasis of this PhD thesis was on examining the intersection between valid self-report of emotion and emotional verbalization and communication, I have demonstrated a lack of external evidence of SSSAC Logistic, suggesting that regulated communication and disconnected communication may better represent emotional communication and verbalization of students engaged in a group work activity. These results suggest two points. First, it would be unwise to interpret SA findings as insights into the emotional state of students. Second, future work should consider how to determine how factors of social coordination and goal orientation influence student communication. Furthermore, rich qualitative explorations of these factors are needed to better understand exactly what SA measures tell us about students.

7.3 UNIQUE CONTRIBUTION

To my knowledge there are no studies in the context of learning which measure the accuracy of SA based on student perceptions where students read text communications from their own group work to identify emotional expression. I reviewed 15 studies (see section 2.1.2.1) in the context of learning and found only three studies which measured the accuracy based on evaluation scores provided by students. When using SA to measure text comments in the evaluation, to the best of my knowledge there are no studies that ask students to (re)read text and provide their judgements about the emotion expression in these texts. When subscribing to a theory of emotion that considers what students perceive to be integral to emotion then this finding

is of even more importance. For example, I considered the CTE where emotion is considered a social construct, where accuracy is at best considered in terms of consensus for students in the social context. From this perspective alignment between SA measures and student perception is of the utmost importance.

I indeed found that using student sourced labels on data produced a higher level of accuracy compared to MTurk labels and general SA benchmarks. This result both demonstrates the benefits of student sourcing labels to train SA classifiers and indicates that in terms of aligning a SA measure with student perceptions existing SA technologies have room for improvement. Based on the CTE these results indicate that existing measures may do a relatively poor job in measuring emotions, providing a strong criticism for using general purpose SA technologies for emotional research in the context of learning. Even if researchers disagree with the CTE, these results do indicate a gap between SA measures and student perceptions. In other words, this PhD indicates that using existing SA measures to directly provide predictions to students would also need to identify a strategy to close the gap between student perceptions and generic SA measures. Effectively these results indicate that either existing SA technologies are not doing a good job of modelling emotion expression in text, or they are at least not ideal measures to directly expose to students, without considering how to bridge the gap between SA and student perception of emotion.

When considering how to support student identifying valence in text I demonstrated that by using ESS the labeled examples provided by students produced higher reliability of student examples. As using emotional sentences starters appears to make it easier for students to identify emotion in text there is reason to consider them as an emotion awareness tool. The literature on computer supported collaborative learning indicated that there are multiple potential benefits for using emotion awareness tools such as increasing engagement (Arguedas et al., 2016; Järvelä & Hadwin, 2013); increasing collaboration (Daradoumis, 2013); increasing self-regulation (Arguedas et al., 2016); improving teachers attitude and feedback (Arguedas et al., 2016); increasing social support and interaction (Daradoumis, 2013; Feidakis et al., 2014); increasing positive emotion after collaboration (Molinari et al., 2016); and increasing transactivity (Molinari et al., 2016);, suggesting that follow-up studies could examine a variety of

questions but the intersection of emotion awareness tools and transactivity might help to explore the extent to which communication represents some form of social regulation.

Finally, the results of this thesis suggested that student emotional communication is likely disjoint from emotional experience. When considering the CTE, these results suggest that regulated communication and misinterpretation of non-emotional communication as emotional are likely contributing factors. I saw evidence of both of these potentially contributing factors to this finding in our sample conversation illustrated in Chapters 4-6. These results both urge researchers to avoid making claims about the internal emotional state of students based on SA and provide the suggestion that future work in SA should focus on exploring how regulation relates to expression. In fact, in a previous study (Hillaire et al., 2018) where students self-report positive emotion and appear to maintain neutral expression as indicated by SA there was a correlation with learning outcomes suggesting SA may be best used in conjunction with self-report emotional measures.

Effectively it might be more interesting to consider how emotion expression is divergent from emotional communication as a means to detect whether students are down regulating or up regulating their emotional expression in terms of the valence of their expression. While SA is likely a poor substitute for self-report of emotion it may provide a means to examine how students are regulating their communication to see which forms of emotion regulation are effective in different learning contexts. This suggestion is actually consistent with the one SA study in our review that found a correlation between emotion expression detected by SA and the learning outcome. In this study (Hillaire et al., 2018) students who self-reported a positive emotional experience who maintained more neutral communication demonstrated more success in answering reading comprehension questions. While previous work has interpreted SA measures in terms of emotion regulation, our unique contribution is to suggest that emotion regulation may play a critical role in interpreting studies that use SA. It might be useful to review all findings from research that uses SA in the context of learning, and reinterpret the findings considering the measure as primarily regulated expression.

7.4 LIMITATIONS

Both Study 1 and Study 2 focused on the validation of a SA classifier which showed promising results. Study 3 focused on correlation considering SA as a potential measure of valid self-report of emotion. First of all, I did not focus on the regulation of emotion in communication and the only regulation measure used, BEQ, did not load on anticipated factors when conducting confirmatory factor analysis. In fact, PANAS also had one item that did not load as expected when conducting confirmatory factor analysis. While I selected both BEQ and PANAS based on evidence of appropriateness the challenges with factor loading suggest that this may have contributed to a lack of evidence for correlation.,

Secondly, another limitation to this work is that it focused on two experiments conducted during a one-hour lab activity. This brief window of measurement may be a contributing factor when considering the lack of evidence for correlation., It is possible that data collected over a longer period of time would have resulted in a different outcome. It is possible that if they gained more familiarity with the platform that the novelty of the environment would not influence their communication. It may take a substantial amount of time for learners to develop a collective intentionality when working together on a group task, in particular when participants are working online. Given that participants were put in an anonymous computer-lab setting (exactly to control some of the mediating factors that may distort dynamic group processes in the “wild”), whereby the only means for communication was the chat function, perhaps some participants felt less inclined to share their positive, negative, neutral, or mixed emotions with participants that they may or may not know. Pragmatically many participants seemed to have just focused on the task (goal) and worked through the World Bank task, without extensively and/or deeply sharing emotions with peers. Although the two SA measurements did pick up substantial number of potential instances of emotional expressions by participants, perhaps these emotions as identified in this thesis may not be as intense as shared and/or perceived by participants. Therefore, future research should focus on how students who work together for a substantial period of time and express their emotion via written text. In particular, it would be extremely relevant to revisit RQ5 and RQ6 and test whether (or not) the

identified patterns by SSSAC Logistic have a stronger relation to self-report measurements of emotions.

Thirdly, this research primarily used a post-positivistic quantitative approach to understand and measure emotions. Although I specifically included the students' perspectives in building the SSSAC, perhaps the actual lived experiences of students might be different. Qualitative approaches that build on the interviews used in Study 1, focus-groups, or critical event recall processes could potentially illicit why some groups of learners expressed more positive, negative, mixed, and neutral emotion relative to others, and whether these quantitative reported differences are actually meaningful.

Nonetheless, I argue that given the size and scale of the three Studies and approaches used these limitations are mitigated. In Study 1 using DBR an initial SSSAC Logistic instrument was developed with 767 students, which is a sufficiently large sample to test a new SA approach that is student-sourced. Subsequently, our RCT in Study 2 confirmed the initial findings with an even larger sample of students, and the intervention of the emotional sentence starter showed results in the expected direction. The results from the replication were subject to four changes in the study detailed in section 5.2.2. Further research will be needed to unpack how SSSAC Logistic is related to other emotion measures, and whether the SSSAC Logistic approach needs to be adjusted for different contexts and time durations when people are working for a longer period of time.

7.5 RECOMMENDATIONS

To evaluate the design of this PhD in terms of what might be logical next steps I break that analysis into three parts. The first part focuses on what might be done differently to create a SA measure for the purposes of finding correlates with emotional measures. The second part emphasizes what could be done differently to pursue correlates with regulated expression. The third part explores what might be needed to consider SA as measuring multiple forms of communication. Each of these directions are reasonable undertakings for different reasons. As I make recommendations for

future work I first explain reasons to pursue one of these options and then interpret the results of this thesis to inform future work in that direction.

7.5.1 FUTURE WORK ON SENTIMENT ANALYSIS FOR VALID SELF-REPORT OF EMOTION

A reason to pursue creating a SA classifier that demonstrates correlates with other emotional measures is that it would be helpful to have an emotional measure that does not require administering a questionnaire to students for two reasons. First, if a SA measure demonstrated correlates with other self-report instruments then I could apply that technology to gain insight into the role of emotion in learning to learning contexts where students have engaged in text-based communication. Second, there are known limitations for self-report and if a SA measure had a high level of convergence with self-report, in places where SA and self-report diverged, SA might simply be accurate in places where students are not forthcoming with providing an accurate self-report. To rigorously examine how to build from this work to achieve a valid self-report of emotion from SA I can step through the process used to generate a SSSAC and consider each decision point.

First, I trained the classifier using a sampling procedure where students selected between 1-3 examples of communications they thought were positive, negative, neutral, and mixed during reflection. I could have asked students to classify their communication when sending messages real-time. A reason to try this is that humans are not perfect historians of their own emotional experience (Pham, 2004) and getting labels from students when they compose their communication may generate labels with higher accuracy. Second, I could use labels from students for their own communication and not include peer labels. I used labels from messages labeled by the author and the intended audience. By selecting just author provided labels this may be closer to the internal psychological state and the measure may demonstrate a higher level of correlates with psychological measures.

7.5.2 FUTURE WORK ON SENTIMENT ANALYSIS FOR REGULATED COMMUNICATION

A reason to pursue correlation testing for measures of regulated communications is that it would directly build on the results of this thesis that there is limited evidence of correlation between SA and psychological measures of emotion. As our theoretical model indicated that communication that was not valid self-report of emotion might be regulated communication it is possible that SA is a measure that better aligns with regulated communication. I examined the trait of emotional expressivity using the BEQ. However, confirmatory factor analysis for the BEQ failed to confirm the expected factors. Future work should consider alternative measure from the BEQ in terms of measuring emotional expressivity. In addition to considering alternative measures it would also be helpful to consider theoretical reasons that students might engage in more regulated communication.

Future studies which focus on correlation between SA and regulated communication may also benefit from using measures for students' goals to see how goal orientation relates to regulated communication. Using the entire BEQ tool might produce better results in terms of confirmatory factor analysis. and it is possible that using a different tool entirely would be better. Our review of the CSCL literature indicated that transactivity (i.e., responding to or building on what a peer has said) is an important aspect to online collaborative learning, so it may be helpful to determine the extent to which communication represents building on what peers are saying as this may be disjoint from internal emotional state and provide a theoretical reason as to why communication is not a valid self-report of emotion.

7.5.3 FUTURE WORK ON SENTIMENT ANALYSIS AND CATEGORIZED COMMUNICATION

It is also possible that SA detects expressions that are both valid self-report of emotion and regulated communication. Testing either in isolation may yield a lack of correlates with other measures. If this is the case then it may be necessary to devise a strategy to classify communications as one or the other before testing correlation., Perhaps when communication is a combination of both there is enough noise when focusing on just one aspect of communication that correlation will fail. Future work may

consider how to classify communication by identifying if students appear to describing their emotional state, or if they are engaging in regulated communication. One potential way forward to doing this might be considering how to script both forms of communication using emotional sentence starters. In this PhD thesis emotional sentence starters appeared to have the desired effect of increasing the accuracy of SA indicating that they are a potential tool to consider when considering measuring emotion expression in text-based communication. Extensions that provide options for regulated communication and self-report communication may help in classification of expressions as either of those categories.

Given that, from our theoretical perspective, communication could be classified as validated self-report of emotion, regulated communication, or disconnected communication I can draw the conclusion that at least enough communication was not validated self-report to prevent correlations with psychometric measures of emotion. So, what is the communication that students produce that expresses emotion and does not align with their internal emotional state and why do they produce it? Our theoretical model suggests that it is either regulated communication or disconnected communication. If students primarily generated regulated communication then why do they explicitly communicate emotion disjoint from their internal state? Are they engaging in some form of social regulation where they are attempting to communicate with their peers using language they anticipate to be accepted and understood over communicating about their internal state? Do they not want their peers to know about their internal state? If they are expressing emotion as a social strategy to achieve some goal then does this potentially indicate that the situated affectivity theory is a better choice to model emotional expression?

7.6 CONCLUSION

There is both promise and peril in using emotional measures in the context of learning. As I have demonstrated in this PhD thesis there is more work necessary to connect SA to a theoretical basis of emotion that is sufficient to measure correlation. As the results of this PhD thesis have not arrived at a clear alignment between a theoretical basis of emotion and SA caution is urged when interpreting SA in the context of learning. Given that there are already promising results from the application of SA to the context of learning I anticipate that more studies will find relationships between SA and learning. It is crucial that those studies avoid making strong claims about students' emotions and measures of SA unless they have more evidence of correlation between emotional expression and internal emotional state than I was able to find with this work. Furthermore, it may be more productive to examine how students' emotional expression relate to their goals or serve as a means of social coordination by focusing on how conscious regulation of emotion results in communication disjoint from internal emotional state.

REFERENCES

- Ainley, M. (2007). Being and Feeling Interested. Transient State, Mood, and Disposition. In *Emotion in Education* (pp. 147–163). <https://doi.org/10.1016/B978-012372545-5/50010-1>
- Alana, B., & Snibbe, C. (2006). Drowning in Data.
- Alonso, O., Marshall, C. C., & Najork, M. (2013). A human-centered framework for ensuring reliability on crowdsourced labeling tasks. *AAAI Workshop - Technical Report, WS-13-18*, 2–3.
- Altrabsheh, N., Gaber, M. M., & Cocca, M. (2013). SA-E : Sentiment Analysis for Education. In *The 5th KES International Conference on Intelligent Decision Technologies (KES-IDT)*. Portugal. <https://doi.org/10.3233/978-1-61499-264-6-353>
- Arguedas, M., Daradoumis, T., & Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation and learning outcome. *Educational Technology and Society*.
- Bagozzi, R. P., Wong, N., & Yi, Y. (1999). The Role of Culture and Gender in the Relationship between Positive and Negative Affect. *Cognition & Emotion*, 13(6), 641–672. <https://doi.org/http://dx.doi.org/10.1080/026999399379023>
- Baker, R. S. J. d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133. Retrieved from <http://eric.ed.gov/?id=ED537205>
- Bakharia, A., Corrin, L., de Barba, P., Kennedy, G., Gasevic, D., Mulder, R., ... Lockyer, L. (2016). A Conceptual Framework linking Learning Design with Learning Analytics Learning Analytics. *6th International Conference on Learning Analytics and Knowledge*. <https://doi.org/10.1145/2883851.2883944>
- Balahur, A., & Steinberger, R. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceedings of the '1st Workshop on Opinion Mining and Sentiment Analysis*, 1–12. Retrieved from http://langtech.jrc.it/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf
- Barchard, K. A., Hensley, S., Anderson, E. D., & Walker, H. E. (2013). Measuring the Ability to Perceive the Emotional Connotations of Written Language. *Journal of Personality Assessment*, 95(4), 332–342. <https://doi.org/10.1080/00223891.2012.736906>
- Barford, K. A., & Smillie, L. D. (2016). Openness and other Big Five traits in relation to dispositional mixed emotions. *Personality and Individual Differences*, 102(November), 118–122. <https://doi.org/10.1016/j.paid.2016.07.002>
- Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 1(1), 28–58. <https://doi.org/10.1111/j.1745-6916.2006.00003.x>
- Barrett, L. F. (2012). Emotions are real. *Emotion*, 12(3), 413–429. <https://doi.org/10.1037/a0027555>
- Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2016). Synthesizing Results From Empirical Research on Computer-Based Scaffolding in STEM Education: A Meta-Analysis. *Review of Educational Research*, 87(2), 1–36.

- <https://doi.org/10.3102/0034654316670999>
- Berrios, R., & Totterdell, P. (2013). Validation of a new Mixed Emotions scale. In *2nd Annual Spring Conference of the White Rose Consortium*.
- Berrios, R., Totterdell, P., & Kellett, S. (2015). Investigating goal conflict as a source of mixed emotions. *Cognition & Emotion*, 29(4), 755–763.
<https://doi.org/10.1177/0146167208325772>
- Blumenthal, A. L. (1975). A reappraisal of Wilhelm Wundt. *American Psychologist*, 30, 1081–1088. <https://doi.org/10.1037/0003-066X.30.11.1081>
- Bosch, N., Mello, S. D., Hall, F., Baker, R., Shute, V., & Wang, L. (2015). Automatic Detection of Learning - Centered Affective States in the Wild.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44(1), 108–132. <https://doi.org/10.1006/jmps.1999.1279>
- Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8(1), 158–233.
<https://doi.org/10.3102/0091732X008001158>
- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, 76(5), 839–855. <https://doi.org/10.1037/0022-3514.76.5.839>
- Cacioppo, J. T., Larsen, J. T., Smith, N. K., & Bertson, G. G. (2004). The affect system: what lurks below the surface of feelings? *Feelings and Emotions The Amsterdam Conference*, (September), 223–242.
- Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37. <https://doi.org/10.1109/T-AFFC.2010.1>
- Calvo, R. A., & Kim, S. Mac. (2010). Sentiment Analysis in Student Experiences of Learning. *Third International Conference on Educational Data Mining (EDM2010)*, 111–120.
- Carrera, P., & Oceja, L. (2017). Drawing mixed emotions : Sequential or simultaneous experiences ? Drawing mixed emotions : Sequential or simultaneous, 9931(January). <https://doi.org/10.1080/02699930600557904>
- CAST. (2018). Universal Design for Learning Guidelines version 2.2.
- Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., & McCrae, J. P. (2020). Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. *ArXiv*.
- Chang, Y. H., Maheswaran, R., Kim, J., & Zhu, L. (2013). Analysis of emotion and engagement in a STEM alternate reality game. *CEUR Workshop Proceedings*, 1009, 29–32. <https://doi.org/10.1007/978-3-642-39112-5-82>
- Chaplot, D. S., Rhim, E., & Kim, J. (2015). Predicting student attrition in MOOCs using sentiment analysis and neural networks. *Workshops at the 17th International Conference on Artificial Intelligence in Education, AIED-WS 2015*, 1432, 7–12.
<https://doi.org/10.1016/j.evalprogplan.2016.04.006>
- Choi, B. (2016). Spring / printemps 2016 How people learn in an asynchronous online learning environment : The relationships between graduate students ' learning strategies and learning satisfaction Comment apprennent les gens dans un environnement d ' apprentissage en li, 42(1), 1–15.

- Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*. *British Journal of Educational Studies* (Vol. 55). https://doi.org/10.1111/j.1467-8527.2007.00388_4.x
- Collignon, O., Girard, S., Gosselin, F., Saint-Amour, D., Lepore, F., & Lassonde, M. (2010). Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia*, 48(1), 220–225. <https://doi.org/10.1016/j.neuropsychologia.2009.09.007>
- Coy, K., Marino, M. T., & Serianni, B. (2014). Using Universal Design for Learning in Synchronous Online Instruction. *Journal of Special Education Technology*, 29(1), 63–74. <https://doi.org/10.1177/016264341402900105>
- Coyne, P., Pisha, B., Dalton, B., Zeph, L. A., & Smith, N. C. (2012). Literacy by Design: A Universal Design for Learning Approach for Students With Significant Intellectual Disabilities. *Remedial and Special Education*, 33(3), 162–172. <https://doi.org/10.1177/0741932510381651>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 49(3), 803–821. <https://doi.org/10.3758/s13428-016-0743-z>
- Crossley, S. A., & McNamara, D. S. (2016). *Adaptive educational technologies for literacy instruction*. *Adaptive Educational Technologies for Literacy Instruction*. Taylor and Francis. <https://doi.org/10.4324/9781315647500>
- Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*, 6–14. <https://doi.org/10.1145/2883851.2883931>
- Crossley, S., Place, P., Mcnamara, D. S., Baker, R. S., & York, N. (2016). Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. *Lak*. <https://doi.org/10.1145/2883851.2883931>
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145–157. <https://doi.org/10.1016/j.learninstruc.2011.10.001>
- D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2012.05.003>
- D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring Affective Trajectories During Complex Learning. In *Proceedings of the 29th Annual Cognitive Science Society* (pp. 203–208). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-1428-6_849
- Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-011-0064-1>
- Daradoumis, T. (2013). Building Intelligent Emotion Awareness for Improving Collaborative e-Learning. <https://doi.org/10.1109/INCoS.2013.49>
- Dawid, A. P., & Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1), 20.

- <https://doi.org/10.2307/2346806>
- de Boer, M. R., Waterlander, W. E., Kuijper, L. D. J., Steenhuis, I. H. M., & Twisk, J. W. R. (2015). Testing for baseline differences in randomized controlled trials: An unhealthy research behavior that is hard to eradicate. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1), 1–8.
<https://doi.org/10.1186/s12966-015-0162-z>
- Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. *Three Worlds of CSCL: Can We Support CSCL?*, 61–91. <https://doi.org/10.1007/s11165-004-8795-y>
- Dinmore, S., & Stokes, J. (2015). Creating inclusive university curriculum: Implementing universal design for learning in an enabling program. *Widening Participation and Lifelong Learning*, 17(4), 4–19.
<https://doi.org/10.5456/WPLL.17.4.4>
- Drake, R. A., Myers, L. R., & Drake, R. (2006). Visual attention, emotion, and action tendency: Feeling active or passive. *Cognition & Emotion*, 20(5), 608–622.
<https://doi.org/10.1080/02699930500368105>
- Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, 44(4), 237–251. <https://doi.org/10.3102/0013189X15584327>
- Edgar, D. W. (2012). Learning Theories and Historical Events Affecting Instructional Design in Education: Recitation Literacy Toward Extraction Literacy Practices. *SAGE Open*, 2(4), 2158244012462707-.
<https://doi.org/10.1177/2158244012462707>
- Ekman, P., Davidson, R., Ellsworth, P., Friesen, W. V., Levenson, R., Oster, H., & Rosenberg, E. (1992). Are There Basic Emotions? *Psychological Review*, 99(3), 550–553.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9(2008), 1871–1874. <https://doi.org/10.1038/oby.2011.351>
- Feidakis, M., Daradoumis, T., Caballé, S., & Conesa, J. (2014). Embedding emotion awareness into e-learning environments. *International Journal of Emerging Technologies in Learning*. <https://doi.org/10.3991/ijet.v9i7.3727>
- Feldman Barrett, L. (2017). *How Emotions are Made*. Macmillon.
- Feldman Barrett, L., & Russell, J. A. (1998). Independence and Bipolarity in the Structure of Current Affect. *Journal of Personality and Social Psychology*, 74(4), 967–984.
- Ferguson, R., & Shum, S. B. (2012). Social Learning Analytics : Five Approaches, (May).
- Fiedler, K., & Beier, S. (2014). Affect and Cognitive Processes in Educational Contexts. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International Handbook of Emotions in Education* (pp. 36–55). New York, New York: Routledge.
- Fleeson, W. (2004). Moving Personality Beyond the Person-Situation Debate: The Challenge and the Opportunity of Within-Person Variability. *Current Directions in Psychological Science*. <https://doi.org/10.1111/j.0963-7214.2004.00280.x>
- Forman, G. (2007). Feature Selection for Text Classification. In H. Liu & H. Motoda (Eds.), *Computational Methods of Feature Selection* (Vol. 16). CRC Press/Taylor

- and Francis Group.
- Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies. *ACM SIGKDD Explorations Newsletter*, 12(1), 49.
<https://doi.org/10.1145/1882471.1882479>
- Freedman, D. H. (2010). Why Scientific Studies Are So Often Wrong : The Streetlight Effect. *Why Scientific Studies Are So Often Wrong : The Streetlight Effect*, (August), 1–8.
- Gallardo-echenique, E., Bullen, M., & Luis Marqués-Molíás. (2016). Student Communication and Study Habits of First-year University Students in the Digital Era Communication étudiante et habitudes d ' étude des étudiants universitaires de première année à l ' époque numérique. *Canadian Journal Of Learning And Technology / La Revue Canadienne De L'Apprentissage Et De La Technologie*, 42(1). Retrieved from <http://cjlt.csj.ualberta.ca/index.php/cjlt/article/view/908>
- Ge, X., & Land, S. M. (2004). A conceptual framework for scaffolding III-structured problem-solving processes using question prompts and peer interactions. *Educational Technology Research and Development*, 52(2), 5–22.
<https://doi.org/10.1007/BF02504836>
- Gill, A. J., Gergle, D., French, R. M., & Oberlander, J. (2008). Emotion rating from short blog texts. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, (January), 1121.
<https://doi.org/10.1145/1357054.1357229>
- Gillioz, C., Gygax, P., & Tapiero, I. (2012). Individual differences and emotional inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 66(4), 239–250.
<https://doi.org/10.1037/a0028625>
- Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement Error Masks Bipolarity in Affect Ratings, 64(6), 1029–1041.
- Gross, J. J., & John, O. P. (1997a). Revealing feelings: facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, 72(2), 435–448. <https://doi.org/10.1037/0022-3514.72.2.435>
- Gross, J. J., & John, O. P. (1997b). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, 72, 435–448. <https://doi.org/10.1017/CBO9781107415324.004>
- Gross, J. J., John, O. P., & Richards, J. M. (2000). The Dissociation of Emotion Expression From Emotion Experience : A Personality Perspective. *Personality and Social Psychology Bulletin*, 26(6), 712–726.
- Grossmann, I., Huynh, A. C., & Ellsworth, P. C. (2015). Emotional Complexity: Clarifying Definitions and Cultural Correlates. *Journal of Personality and Social Psychology*, 1(DECEMBER). <https://doi.org/10.1017/CBO9781107415324.004>
- Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., & Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition and Learning*, 2(2–3), 107–124. <https://doi.org/10.1007/s11409-007-9016-7>
- Hall, T. E., & Vue, G. (2012). Transforming Writing Instruction with Universal Design for Learning. In T. E. Hall, A. Meyer, & D. H. Rose (Eds.), *Universal Design for Learning in the Classroom Practical Applications* (pp. 38–54). Guilford Press.
- Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An

- Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, (April), 929–932. <https://doi.org/10.1145/1240624.1240764>
- Hershfield, H. E., & Larsen, J. T. (2012). On the measurement of mixed emotions: A critical review. *White Paper Commissioned by the National Institute on Aging*.
- Hillaire, G. (2015). Emotional Experiences of Students. In *Open University Postgraduate Research Student Poster Competition*. Milton Keynes, UK. <https://doi.org/10.13140/RG.2.2.26113.76640>
- Hillaire, G., Fenton-O'Creevy, M., Mittelmeier, J., Rienties, B., Tempelaar, D., & Zdrahal, Z. (n.d.). *A student-sourced sentiment analysis to improve detection of (mixed) emotion in online discourse*.
- Hillaire, G., Rappolt-Schlichtmann, G., & Ducharme, K. (2016). Prototyping Visual Learning Analytics Guided by an Educational Theory Informed Goal. Garron.Hillaire@open.ac.uk. *Journal of Learning Analytics*, 3(3), 115–142. <https://doi.org/10.18608/jla.2016.33.7>
- Hillaire, G., Rappolt-Schlichtmann, G., & Stahl, W. (2014). Learning Analytics Summer Institute 2014 – Panel – Universal Design for Learning. Retrieved from <https://youtu.be/Sl-6QATSw3E>
- Hillaire, G., Rienties, B., & Goldowsky, B. (2018). Struggling Readers Smiling on the Inside and Getting Correct Answers. In *AERA 2018 Annual Meeting, Symposium Session: Perspectives on Emotion and Engagement Among Struggling Adolescent Readers: Findings from an Online Literacy Intervention*. New York, New York.
- Hsueh, P.-Y., Melville, P., & Sindhvani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, (June), 27–35. <https://doi.org/10.1.1.157.5154>
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*, 168. <https://doi.org/10.1145/1014052.1014073>
- Hui, C. M., Fok, H. K., & Bond, M. H. (2009). Who feels more ambivalence? Linking dialectical thinking to mixed emotions. *Personality and Individual Differences*, 46(4), 493–498. <https://doi.org/10.1016/j.paid.2008.11.022>
- Hupont, I., Lebreton, P., Maki, T., Skodras, E., & Hirth, M. (2014). Is affective crowdsourcing reliable? *2014 IEEE 5th International Conference on Communications and Electronics, IEEE ICCE 2014*, (July), 516–521. <https://doi.org/10.1109/CCE.2014.6916757>
- Hutchison, D., & Styles, B. (2010). A guide to running randomised controlled trials for educational researchers, 1–65.
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and ...*, 216–225. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109%5Cnhttp://comp.social.gatech.edu/papers/icwsml4.vader.hutto.pdf>

- Immordino-Yang, M. H. (2016). *Emotions, Learning, and the Brain*. W. W. Norton & Co.
- Immordino-Yang, M. H., & Damasio, A. (2007). We Feel, Therefore We Learn: The Relevance of Affective and Social Neuroscience to Education. *Mind, Brain, and Education*, 1(1), 3–10. <https://doi.org/10.1111/j.1751-228X.2007.00004.x>
- Iqbal, M., Karim, A., & Kamiran, F. (2015). Bias-Aware Lexicon-Based Sentiment Analysis. In *SAC '15 Proceedings of the 30th Annual ACM Symposium on Applied Computing* (pp. 845–850). <https://doi.org/http://dx.doi.org/10.1145/2695664.2695759>
- Jagtap, B., & Dhotre, V. (2014). SVM and HMM Based Hybrid Approach of Sentiment Analysis for Teacher Feedback Assessment. *Ijettcs.Org*, 3(3), 229–232. Retrieved from <http://ijettcs.org/Volume3Issue3/IJETTCS-2014-06-25-132.pdf>
- Järvelä, S. (2014). Regulated Learning in CSCL : Theoretical Progress for Learning Success Socially shared regulation of learning in CSCL : Understanding and prompting individual and group level shared regulatory activities, (c).
- Järvelä, S., & Hadwin, A. F. (2013). New Frontiers: Regulating Learning in CSCL. *EDUCATIONAL PSYCHOLOGIST*, 48(1), 25–39. <https://doi.org/10.1080/00461520.2012.748006>
- Järvelä, S., Järvenoja, H., Malmberg, J., & Hadwin, A. F. (2013). Exploring socially shared regulation in the context of collaboration. *Journal of Cognitive Education and Psychology*, 12(3), 267–286. <https://doi.org/10.1891/1945-8959.12.3.267>
- Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., & Verykios, V. S. (2015). A Learning Analytics Methodology for Detecting Sentiment in Student Fora: A Case Study in Distance Education. *European Journal of Open, Distance and E-Learning*, 18(2). <https://doi.org/10.1515/eurodl-2015-0014>
- Kahn, J. H., Tobin, R. M., Massey, A. E., Anderson, J. A., The, S., Journal, A., & Summer, N. (2016). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, 120(2), 263–286.
- Kalpana, T. (2014). A Constructivist Perspective on Teaching and Learning: A Conceptual Framework. *International Research Journal of Social Sciences*, 3(1), 27–29. Retrieved from <http://www.isca.in/IJSS/Archive/v3/i1/6.ISCA-IRJSS-2013-186.pdf>
- Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional Expression : Advances in Basic Emotion Theory. *Journal of Nonverbal Behavior*, (0123456789). <https://doi.org/10.1007/s10919-019-00293-3>
- Khetan, A., Lipton, Z. C., & Anandkumar, A. (2017). Learning from noisy singly-labeled data. *ArXiv*, (1979), 1–15.
- Knight, S., Rienties, B., Littleton, K., Mitsui, M., Tempelaar, D., & Shah, C. (2017). The relationship of (perceived) epistemic cognition to interaction with resources on the internet. *Computers in Human Behavior*, 73, 507–518. <https://doi.org/10.1016/j.chb.2017.04.014>
- Koehler, M. J., Greenhalgh, S., & Zellner, A. (2015). Potential Applications of Sentiment Analysis in Educational Research and Practice – Is SITE the Friendliest Conference? D. Rutledge & D. Slykhuis (Eds.), *Proceedings of SITE 2015--Society for Information Technology & Teacher Education International Conference*, 1348–1354. Retrieved from

- <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.859.5705&rep=rep1&type=pdf>
- Kreibig, S. D., & Gross, J. J. (2017). Understanding mixed emotions: paradigms and measures. *Current Opinion in Behavioral Sciences*, 15, 62–71.
<https://doi.org/10.1016/j.cobeha.2017.05.016>
- Krippendorff, K. (1980). *Analysis: An Introduction to its Methodology*.
<https://doi.org/10.2307/2288384>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Lang, C., Siemens, G., Wise, A. F., & Gasevic, D. (2017). *Handbook of Learning Analytics*. (C. Lang, G. Siemens, A. Wise, & D. Gasevic, Eds.). Society for Learning Analytics Research (SoLAR). <https://doi.org/10.18608/hla17>
- Larsen, J. T., Coles, N. A., & Jordan, D. K. (2017). Varieties of mixed emotional experience. *Current Opinion in Behavioral Sciences*, 15, 72–76.
<https://doi.org/10.1016/j.cobeha.2017.05.021>
- Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology*, 81(4), 684–696.
<https://doi.org/10.1037//0022-3514.81.4.684>
- Lawrence, C., & Rodriguez, P. (2012). The interpretation and legitimization of values in agile's organizing vision. *Ecis*, (2012). Retrieved from
<http://www.scopus.com/inward/record.url?eid=2-s2.0-84905734781&partnerID=tZOtx3y1>
- Lerner, J. S., & Keltner, D. (2015). Cognition and Emotion Beyond valence: Toward a model of emotion- specific influences on judgement and choice Beyond valence: Toward a model of emotion-speci® c in¯ uences on judgement and choice. *Cognition & Emotion*, 14(4), 473–493. <https://doi.org/10.1080/026999300402763>
- Leue, A., & Beauducel, A. (2011). The PANAS Structure Revisited: On the Validity of a Bifactor Model in Community and Forensic Samples. *Psychological Assessment*, 23(1), 215–225. <https://doi.org/10.1037/a0021400>
- Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*.
- Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2012). The Hundred-Year Emotion War: Are Emotions Natural Kinds or Psychological Constructions? Comment on Lench, Flores, and Bench (2011), 40(6), 1301–1315.
<https://doi.org/10.1007/s10439-011-0452-9.Engineering>
- Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience*, 6(5), 1–6.
<https://doi.org/10.1093/gigascience/gix020>
- Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, (1), 1–38. <https://doi.org/10.1145/1772690.1772756>
- Liu, B., & Street, S. M. (2005). Opinion Observer : Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th International Conference on World Wide Web*, 342–351. <https://doi.org/10.1145/1060745.1060797>
- Lonchamp, J. (2012). Computational analysis and mapping of ijCSCL content, (February), 475–497. <https://doi.org/10.1007/s11412-012-9154-z>

- Ludvigsen, S. (2016). CSCL : connecting the social , emotional and cognitive dimensions. *International Journal of Computer-Supported Collaborative Learning*, (1), 115–121. <https://doi.org/10.1007/s11412-016-9236-4>
- Malatesta, C. Z., & Haviland, J. M. (1982). Leammg Display Rules: The Socialization of Emotion Expression in Infancy. *Child Development*, 53(99), 991–1003.
- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, 23(2), 209–237. <https://doi.org/10.1080/02699930802204677>
- Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? *Emotional Development and Emotional Intelligence*. <https://doi.org/10.1177/1066480710387486>
- McKenney, S., & Reeves, T. C. (2014). Educational Design Research. In *Handbook of Research on Educational Communications and Technology* (pp. 131–140). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_11
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning Theory and Practice*. Wakefield, MA: Cast Incorporated.
- Mittelmeier, J., Rienties, B., Tempelaar, D., Hillaire, G., & Whitelock, D. (2018). The influence of internationalised versus local content on online intercultural collaboration in groups: A randomised control trial study in a statistics course. *Computers and Education*, 118, 82–95. <https://doi.org/10.1016/j.compedu.2017.11.003>
- Mittelmeier, J., Rienties, B., Tempelaar, D., & Whitelock, D. (2017). Overcoming cross-cultural group work tensions: mixed student perspectives on the role of social relationships. *Higher Education*, 1–18. <https://doi.org/10.1007/s10734-017-0131-3>
- Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, (June), 26–34. Retrieved from <http://dl.acm.org/citation.cfm?id=1860631.1860635>
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. In *Computational Intelligence*. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Molinari, G., Chanel, G., Bétrancourt, M., Pun, T., & Bozelle, C. (2016). Emotion feedback during computer-mediated collaboration: Effects on self-reported emotions and perceived interaction. *Computer-Supported Collaborative Learning Conference, CSCL, 1(2009)*, 336–343. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-84886502061&partnerID=40&md5=d97d49119837695d415c2e8cfb3a1a6c>
- Morris, R., Hadwin, A. F., Gress, C. L. Z., Miller, M., Fior, M., Church, H., & Winne, P. H. (2010). Designing roles, scripts, and prompts to support CSCL in gStudy. *Computers in Human Behavior*, 26(5), 815–824. <https://doi.org/10.1016/j.chb.2008.12.001>
- Morris, R., & McDuff, D. (2009). Crowdsourcing Techniques for Affective Computing. In *Handbook of Affective Computing* (Vol. 14, pp. 27–35).

- <https://doi.org/10.1177/1354856507084420>
- Mosier, C. I. (1941). A Psychometric Study of Meaning. *Journal of Social Psychology*, 13(1), 123–140. <https://doi.org/10.1080/00224545.1941.9714065>
- Muller, M., Wolf, C. T., Andres, J., Ashktorab, Z., Narendra, N. J., Desmond, M., ... Dugan, C. (2021). Designing Ground Truth and the Social Life of Labels ACM Reference Format Designing Ground Truth and the Social Life of Labels. In *CHI2021*. Yokohama, Japan. <https://doi.org/10.1145/11445.11456>
- Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/TAFFC.2014.2317187>
- Munezero, M., Mozgovoy, M., Montero, C. S., & Sutinen, E. (2013). Exploiting Sentiment Analysis to Track Emotions in Students' Learning Diaries. In *Koli Calling*. <https://doi.org/10.1145/2526984>
- Näykki, P., Järvelä, S., Kirschner, P. A., & Järvenoja, H. (2014). Socio-emotional conflict in collaborative learning-A process-oriented case study in a higher education context. *International Journal of Educational Research*, 68, 1–14. <https://doi.org/10.1016/j.ijer.2014.07.001>
- Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly*, 46(3), 273–304. <https://doi.org/10.1598/RRQ.46.3.4>
- Noroozi, O., Teasley, S. D., Biemans, H. J. A., Weinberger, A., & Mulder, M. (2013). *Facilitating learning in multidisciplinary groups with transactive CSCL scripts. International Journal of Computer-Supported Collaborative Learning* (Vol. 8). <https://doi.org/10.1007/s11412-012-9162-z>
- Nussbaum, E. M., Hartley, K., Sinatra, G. M., Reynolds, R. E., & Bendixen, L. D. (2005). Personality Interactions and Scaffolding in On-Line Discussions. *Journal of Educational Computing Research*, 30(1–2), 113–137. <https://doi.org/10.2190/h8p4-qjuf-jxme-6jd8>
- Okon-Singer, H., Hendler, T., Pessoa, L., & Shackman, A. J. (2015). The neurobiology of emotion cognition interactions: fundamental questions and strategies for future research. *Frontiers in Human Neuroscience*, 9(February), 58. <https://doi.org/10.3389/fnhum.2015.00058>
- Op ' , P., Eynde, T., De Corte, E., & Verschaffel, L. (2007). Students' Emotions: A Key Component of Self-Regulated Learning? In *Emotion in Education* (pp. 185–204).
- Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31, 527–541. <https://doi.org/10.1016/j.chb.2013.05.024>
- Pang, B., & Lee, L. (2006). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Informatio*Pang, B., & Lee, L. (2006). *Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval*, 1(2), 91–231. *Doi:10.1561/1500000000In Retrieval*, 1(2), 91–231. <https://doi.org/10.1561/1500000001>
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Empirical Methods in Natural Language Processing (EMNLP)*, 10(July), 79–86. <https://doi.org/10.3115/1118693.1118704>

- Pardos, Z. A., Baker, R. S. J., & Pedro, M. O. C. Z. S. (2013). Affective states and state tests : Investigating how affect throughout the school year predicts end of year learning outcomes, 117–124.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Pekrun, R. (2005). Progress and open problems in educational emotion research. *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2005.07.014>
- Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, 36(1), 36–48. <https://doi.org/10.1016/j.cedpsych.2010.10.002>
- Perry, N. E., & Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, 18(3), 211–228. <https://doi.org/10.1007/s10648-006-9014-3>
- Peterson, E. R., Brown, G. T. L., & Jun, M. C. (2015). Achievement emotions in higher education: A diary study exploring emotions across an assessment event. *Contemporary Educational Psychology*, 42, 82–96. <https://doi.org/10.1016/j.cedpsych.2015.05.002>
- Pham, M. T. (2004). The Logic of Feeling. *Journal of Consumer Psychology*, 14(4), 360–369. https://doi.org/10.1207/s15327663jcp1404_5
- Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., ... Strohecker, C. (2004). Affective Learning — A Manifesto. *BT Technology Journal*, 22(4), 253–269. <https://doi.org/10.1023/B:BTJ.0000047603.37042.33>
- Pintrich, P. R., & De Groot, E. V. (1990). Motivational and Self-Regulated Learning Components of Classroom Academic Performance. *Journal of Educational Psychology*, 82(1), 33–40. <https://doi.org/10.1037/0022-0663.82.1.33>
- Posey, A. (2018). How to Support the Emotional Link to Learning. Retrieved from <http://www.ascd.org/ascd-express/vol13/1319-posey.aspx>
- Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-Based Sentiment Analysis of Teachers' Evaluation. *Applied Computational Intelligence and Soft Computing*, 2016, 1–12. <https://doi.org/10.1155/2016/2385429>
- Rappolt-Schlichtmann, G., & Daley, S. G. (2013). Providing Access to Engagement in Learning: The Potential of Universal Design for Learning in Museum Design. *Curator: The Museum Journal*, 56(3), 307–321. <https://doi.org/10.1111/cura.12030>
- Raykar, V. C., Yu, S., Zhao, L. H., Hermosillo Valadez, G., Florin, C., Bogoni, L., ... Org, L. M. (2010). Learning From Crowds. *Journal of Machine Learning Research*, 11, 1297–1322. <https://doi.org/10.2139/ssrn.936771>
- Reeck, C., Ames, D. R., & Ochsner, K. N. (2016). The Social Regulation of Emotion: An Integrative, Cross-Disciplinary Model. *Trends in Cognitive Sciences*, 20(1), 47–63. <https://doi.org/10.1016/j.tics.2015.09.003>
- Rienties, B., Giesbers, B., Tempelaar, D., Lygo-Baker, S., Segers, M., & Gijssels, W. (2012). The role of scaffolding and motivation in CSCL. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2012.04.010>
- Rienties, B., Nguyen, Q., Holmes, W., & Reedy, K. (2017). A review of ten years of

- implementation and research in aligning learning design with learning analytics at the Open University UK. *Interaction Design and Architecture(s) Journal*, 33(Ld), 134–154. Retrieved from http://www.mifav.uniroma2.it/inevent/events/idea2010/doc/33_7.pdf
- Rienties, B., & Rivers, B. A. (2014). Measuring and Understanding Learner Emotions : Evidence and Prospects. *Lace*, 1–16. Retrieved from <http://www.laceproject.eu/publications/learning-analytics-and-emotions.pdf>
- Rienties, B., & Toetenel, L. (2016). The Impact of 151 Learning Designs on Student Satisfaction and Performance: Social Learning (Analytics) Matters, 339–343. <https://doi.org/10.1145/2883851.2883875>
- Roan, L., Strong, B., Foss, P., Yager, M., Gehlbach, H., & Metcalf, K. A. (2009). Social Perspective Taking.
- Robinson, K. (2013). The interrelationship of emotion and cognition when students undertake collaborative group work online: An interdisciplinary approach. *Computers and Education*. <https://doi.org/10.1016/j.compedu.2012.11.003>
- Rose, D. H. (2015). Universal Design in Education Conference Keynote Presentation: From Affect to Effect - Why emotional design is at the core of Universal Design for Learning. Retrieved from <https://www.youtube.com/watch?v=shsfhDqZ1ss&feature=youtu.be>
- Rose, D. H., Harbour, W. S., Johnston, C. S., Daley, S. G., & Abarbanell, L. (2006). Universal design for learning in postsecondary education: Reflections on principles and their application. *Journal of Postsecondary Education and Disability*, 19(2), 135–151.
- Ruiz, S., Charleer, S., Fernández-castro, I., & Duval, E. (2016). Supporting learning by considering emotions : Tracking and Visualization . A case study. <https://doi.org/10.1145/2883851.2883888>
- Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Russell, J. A. (1983). Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45(6), 1281–1288. <https://doi.org/10.1037/0022-3514.45.6.1281>
- Russell, J. A. (1991). Culture and the Categorization of Emotions. *Psychological Bulletin*, 110(3), 426–450.
- Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion, 110(1), 145–172. <https://doi.org/10.1037/0033-295X.110.1.145>
- Russell, J. A., & Barrett, L. F. (1999a). Core Affect , Prototypical Emotional Episodes , and Other Things Called Emotion : Dissecting the Elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819.
- Russell, J. A., & Barrett, L. F. (1999b). Core Affect , Prototypical Emotional Episodes , and Other Things Called Emotion : Dissecting the Elephant. *Journal of Personality and Social Psychology*, 76(5). <https://doi.org/10.1037/0022-3514.76.5.805>
- Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, 125(1), 3–30. <https://doi.org/10.1037/0033-2909.125.1.3>
- Salminen, J. O., Al-Merekhi, H. A., Dey, P., & Jansen, B. J. (2018). Inter-Rater Agreement for Social Computing Studies. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (Vol. 49, pp.

- 80–87). IEEE. <https://doi.org/10.1109/SNAMS.2018.8554744>
- Santos, O. C., Salmeron-Majadas, S., & Boticario, J. G. (2013). Emotions detection from math exercises by combining several data sources. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7926 LNAI, pp. 742–745). <https://doi.org/10.1007/978-3-642-39112-5-102>
- Schachter, S., & Singer, J. E. (1962). COGNITIVE, SOCIAL, AND PHYSIOLOGICAL DETERMINANTS OF EMOTIONAL STATE. *American Psychologist*, 69(5), 379–399. <https://doi.org/10.1037/h0021465>
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. <https://doi.org/10.1177/0539018405058216>
- Scherer, K. R. (2009). Emotions are emergent processes: They require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3459–3474. <https://doi.org/10.1098/rstb.2009.0141>
- Schmidt, T., & Burghardt, M. (2018). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, (2005), 139–149.
- Shao, B., Doucet, L., & Caruso, D. R. (2014). Universality Versus Cultural Specificity of Three Emotion Domains: Some Evidence Based on the Cascading Model of Emotional Intelligence. *Journal of Cross-Cultural Psychology*, 46(2), 229–251. <https://doi.org/10.1177/0022022114557479>
- Shapiro, H. B., Lee, C. H., Wyman, N. E., Li, K., Çetinkaya-rundel, M., & Canelas, D. A. (2017). Understanding the massive open online course (MOOC) student experience : An examination of attitudes , motivations , and barriers. *Computers & Education*, 110, 35–50. <https://doi.org/10.1016/j.compedu.2017.03.003>
- Shawe-Taylor, J., & Cristianini, N. (2000). Support Vector Machines, 1, 1–8. <https://doi.org/10.1007/978-0-387-77242-4>
- Shickel, B., Heesacker, M., Benton, S., Ebadi, A., Nickerson, P., & Rashidi, P. (2016). Self-Reflective Sentiment Analysis, 23–32.
- Shum, S. B., & Crick, R. D. (2012). Learning Dispositions and Transferable Competencies: Pedagogy, Modelling and Learning Analytics. *2nd International Conference on Learning Analytics Knowledge 29 Apr 02 May 2012*, (May). <https://doi.org/10.4018/978-1-4666-0300-4.ch017>
- Siepmann, T., Spieth, P. M., Kubasch, A. S., Penzlin, A. I., Illigens, B. M.-W., & Barlinn, K. (2016). Randomized controlled trials - a matter of design. *Neuropsychiatric Disease and Treatment*, 12, 1341. <https://doi.org/10.2147/NDT.S101938>
- Stapor, K. (2017). Evaluating and Comparing Classifiers: Review, Some Recommendations and Limitations. In *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017* (Vol. 578, pp. 12–21). <https://doi.org/10.1007/978-3-319-59162-9>
- Stinson, L. (2016). Facebook reactions, the totally redesigned like button, is here. Retrieved August 5, 2016, from http://www.wired.com/2016/02/facebook-reactions-totally-redesigned-like-button/?mbid=nl_22416
- Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most

- informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167.
<https://doi.org/10.1016/j.chb.2014.05.038>
- Tempelaar, D. T., Rienties, B., & Nguyen, Q. (2017). Towards Actionable Learning Analytics Using Dispositions, 10(1), 6–16.
- Thelwall, M. (2013). Sentiment Strength Detection for the Social Web. Retrieved from <http://sentistrength.wlv.ac.uk/documentation/course.html>
- Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42(3), 343–354. <https://doi.org/10.1108/OIR-05-2017-0153>
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558.
<https://doi.org/10.1002/asi.21416>
- Tracy, J. L., & Randles, D. (2011). Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, 3(4), 397–405. <https://doi.org/10.1177/1754073911410747>
- Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013). Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning. *IISA 2013 - 4th International Conference on Information, Intelligence, Systems and Applications*, 198–205. <https://doi.org/10.1109/IISA.2013.6623713>
- Turney, P. D. (2002). Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 417–424). Philadelphia, Pennsylvania. <https://doi.org/10.3115/1073083.1073153>
- Twitter Emotion Coding Instructions. (2013), 1(2005), 1–6. Retrieved from <http://sentistrength.wlv.ac.uk/documentation/TwitterVersionOfSentimentCodeBook.doc>
- Uzun, A. M., & Zahide, Y. (2018). Exploring the effect of using different levels of emotional design features in multimedia science learning. *Computers and Education*, 119(January), 112–128. <https://doi.org/10.1016/j.compedu.2018.01.002>
- Van Den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and Cognitive Factors Driving Teamwork in Collaborative Learning Environments Team Learning Beliefs and Behaviors. *Small Group Research*, 37(5), 490–521.
<https://doi.org/10.1177/1046496406292938>
- Västfjäll, D., & Gärling, T. (2007). Validation of a Swedish short self-report measure of core affect. *Scandinavian Journal of Psychology*, 48, 233–238.
<https://doi.org/10.1111/j.1467-9450.2007.00595.x>
- Vygotsky, L. (1978). *Mind in Society*. (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, Massachusetts: Harvard University Press.
- Waldinger, R. J., Hauser, S. T., Schulz, M. S., Allen, J. P., & Crowell, J. A. (2004). Reading others emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology : JFP : Journal of the Division of Family Psychology of the American Psychological Association (Division 43)*, 18(1), 58–71. <https://doi.org/10.1037/0893-3200.18.1.58>
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–

1207. <https://doi.org/10.3758/s13428-012-0314-x>
- Watson, D., Clark, L. a, & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two activation systems of affect: Structural findings, evolutionary considerations, and psychbiological evidence. *Journal of Personality and Social Psychology*, 76(5), 820–838. [https://doi.org/0022-3514/99/\\$3.00](https://doi.org/0022-3514/99/$3.00)
- Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2016). The Jingle and Jangle of Emotion Assessment: Imprecise Measurement, Casual Scale Usage, and Conceptual Fuzziness in Emotion Research. *Emotion*, 17(2), 267–295. <https://doi.org/10.1037/emo0000226>
- Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning. *Instructional Science*, 33(1), 1–30. <https://doi.org/10.1007/s11251-004-2322-4>
- Weinberger, A., Stegmann, K., Fischer, F., & Mandl, H. (1997). Chapter 12 SCRIPTING ARGUMENTATIVE KNOWLEDGE CONSTRUCTION IN COMPUTER-SUPPORTED.
- Wen, M., Yang, D., & Rosé, C. P. (2014). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? *Proceedings of Educational Data Mining*, (Edm), 1–8. Retrieved from <http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf>
- Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3), 165–210. <https://doi.org/10.1007/s10579-005-7880-9>
- Wilson, T., Wiebe, J., & Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. *Acl*, 7(5), 12–21. <https://doi.org/10.3115/1220575.1220619>
- Wilutzky, W. (2015). Emotions as pragmatic and epistemic actions. *Frontiers in Psychology*, 6(OCT), 1–10. <https://doi.org/10.3389/fpsyg.2015.01593>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as Self-Regulated Learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition, Self-Regulated Learning, and Study Strategies* (pp. 277–304). Mahwah, NJ: Erlbaum.
- Wise, A. F., & Vytasek, J. (2017). Learning Analytics Implementation Design. *Handbook of Learning Analytics*, 151–160. <https://doi.org/10.18608/hla17.013>
- Wyner, S., Shaw, E., Kim, T., Li, J., & Kim, J. (2008). Sentiment analysis of a student Q&A board for computer science. *The 9th KOCSEA Technical Symposium*.
- Xia, R., Xu, F., Yu, J., Qi, Y., & Cambria, E. (2016). Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. *Information Processing and Management*. <https://doi.org/10.1016/j.ipm.2015.04.003>
- Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, 50(2), 1–33. <https://doi.org/10.1145/3057270>
- Yik, M., Russell, J. A., & Steiger, J. H. (n.d.). A 12-Point Circumplex Structure of Core

- Affect. <https://doi.org/10.1037/a0023980>
- Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2016). Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5), 541–552. <https://doi.org/10.14778/3055540.3055547>
- Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*. <https://doi.org/10.1207/s15326985ep2501>
- Zimmerman, B. J., Heart, N., Mellins, R. B., & Zimmerman, B. J. (1989). A Social Cognitive View of Self-Regulated Academic Learning. *Journal of Educational Psychology*, 81(3), 329–339. <https://doi.org/10.1037//0022-0663.81.3.329>
- Zimmerman, B. J., & Martinez-pons, M. (1990). Student Differences in Self-Regulated Learning : Relating Grade , Sex , and Giftedness to Self-Efficacy and Strategy Use, 82(1), 51–59.

APPENDICES

APPENDIX 1 – BEQ

1. THE ENTIRE TOOL

Scale (take directly from <http://psychology.stanford.edu/~psyphy/resources.html>):
For each statement below, please indicate your agreement or disagreement. Do so by filling in the blank in front of each item with the appropriate number from the following rating scale:

1-----2-----3-----4-----5-----6-----7
1 = Strongly disagree
4 = Neutral
7 = Strongly agree

1. Whenever I feel positive emotions, people can easily see exactly what I am feeling.
2. I sometimes cry during sad movies.
3. People often do not know what I am feeling.
4. I laugh out loud when someone tells me a joke that I think is funny.
5. It is difficult for me to hide my fear.
6. When I'm happy, my feelings show.
7. My body reacts very strongly to emotional situations.
8. I've learned it is better to suppress my anger than to show it.
9. No matter how nervous or upset I am, I tend to keep a calm exterior.
10. I am an emotionally expressive person.
11. I have strong emotions.
12. I am sometimes unable to hide my feelings, even though I would like to.
13. Whenever I feel negative emotions, people can easily see exactly what I am feeling.
14. There have been times when I have not been able to stop crying even though I tried to stop.
15. I experience my emotions very strongly.
16. What I'm feeling is written all over my face.

2. ITEMS ADMINISTERED

Subset of items selected with partner teacher at site

- Whenever I feel positive emotions, people can easily see exactly what I am feeling.
- People often do not know what I am feeling.
- I have strong emotions.
- When I'm happy, my feelings show.
- No matter how nervous or upset I am, I tend to keep a calm exterior.
- I am sometimes unable to hide my feelings, even though I would like to.

- I am an emotionally expressive person.
- Whenever I feel negative emotions, people can easily see exactly what I am feeling.
- I experience my emotions very strongly.

APPENDIX 2 – PANAS

1. THE ENTIRE TOOL ADMINISTERED

In the post lab survey section students were asked to: describe your group work experience

1-----2-----3-----4-----5
 1 = Not at all
 5 = Extremely

Please rate from "Very Slightly or Not At All" to "Extremely"

- Guilt
- Fear
- Sadness
- Hostility
- Shyness
- Fatigue
- Serenity
- Attentiveness
- Self-Assurance
- Joviality
- Surprise

APPENDIX 3 – MES

1. THE ENTIRE TOOL

“Mixed Emotions scale, this scale measure the presence of mixed feelings regarding to an important event or experience in the last week, using a 5-points scale ranging from “not at all” to “very much”” (Berrios & Totterdell, 2013)

1-----2-----3-----4-----5

1 = Not at all
5 = Very Much

1. I felt a mixture of emotions.
2. I felt a combination of different emotions at the time.
3. I felt different emotions at the same time.
4. I felt contrasting emotions.
5. I felt as if positive emotions and negative emotions had been fused into one feeling.
6. I felt one emotion immediately followed by another emotion.
7. I felt different emotions occur very quickly one after another.
8. I felt a kind of bittersweet feeling.
9. I felt something neither good nor bad, but certainly a truly emotional experience.
10. I felt only one thing throughout the event or experience*.
11. I felt just one emotion very clearly*.
12. I felt either positive or negative emotions but not both at the same time*.
13. I think it is useful to feel mixed emotions.

* Reverse Coded

2. ITEMS ADMINISTERED

In the post lab survey section students were asked to: describe your group work experience

1-----2-----3-----4-----5

1 = Not at all
5 = Very Much

1. I felt a mixture of emotions.
2. I felt different emotions occur very quickly one after another.
3. I felt only one thing throughout the event or experience*.
4. I felt either positive or negative emotions but not both at the same time*.

* Reverse Coded

APPENDIX 4 – POST ACTIVITY FOR EXPERIMENT 1

This was co-designed with Jenna Mittelmeier so some items are in support of her research interests while Section 4 represents my unique contribution in support of this PhD thesis work.

Thank you for participating in this post-activity.

You recently collaborated with a small group of your classmates in a computer lab on a case study related to education statistics. The following questions and activities is the second requirement for receiving your participation points for this week's assignment.

Throughout the activity, you will be asked to reflect upon the group work process and soft skills required to work with diverse group members. At the end of this session, you will be given the opportunity to provide constructive feedback about your group members' contributions to the activity.

In return, the feedback provided by your group members about your own participation in this assignment will be provided to your university email address in approximately two weeks. We hope that this feedback will serve as a useful tool for further developing essential soft skills and cross-cultural competencies for your future career.

Name

Question [short answer]	Name:
-----------------------------------	-------

StudentID

Question [short answer]	Student ID number:
-----------------------------------	--------------------

[Next Page]

SECTION 1 - REFLECTION AND RECALL OF GROUP WORK PROCESS

For the first section of this post-assignment, we would like you to reflect on the process of working together and the final outcome of your group assignment.

Section1_Q1 – Open Ended

Question [Open-ended]	In approximately 200 words, please reflect on the processes and outcomes of working together with your classmates during this computer lab assignment:
---------------------------------	--

Section1_Q2 – Likert (4 items)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	Please rank your agreement to the following statements: 1. We have completed the task in a way we all agree upon.
---	--

	<ol style="list-style-type: none"> 2. I am not satisfied with the performance of our group. 3. I would wish to work with this group in the future 4. As a group, we have learned a lot.
--	--

[Next Page]

During this lab activity, you had the opportunity to work with a diverse group of students from different countries. In the next questions, we would like you to reflect on the composition of your group members and how it affected your ability to work together to complete this assignment.

Section1_Q3 - Open Ended

Question [open-ended]	In what ways was working with diverse group members a benefit to you and your group?
---------------------------------	--

Section1_Q4 – Open Ended

Question [open-ended]	In what ways did working with diverse group members in this assignment lead to difficulties or tensions for you or your group members?
---------------------------------	--

[Next Page]

SECTION 2 - REFLECTION OF ASSIGNMENT CONTENT

We would now like you to consider the content and topic of your assigned case study. You can view this again to refresh yourself by logging on to Udio at: <http://iet-projects.open.ac.uk/cethub>

Your Udio username is your first and last name (example: JohnSmith). The username is case sensitive.

Your password is your birthdate. For example, if you were born on July 1, 1996, your password would be 07011996.

Please consider how the content of the assigned case study influenced your group's interaction with one another:

Section2_Q1 – likert (4 items)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	Please rank your agreement to the following statements: <ol style="list-style-type: none"> 1. I found the assignment intellectually challenging and stimulating (Section2_Q1A) 2. I did not learn something which I consider valuable in this assignment (Section2_Q1B) 3. My interest in the subject has increased as a consequence of this assignment (Section2_Q1C) 4. I have learned and understood the subject matter of this assignment (Section2_Q1D)
---	--

Your case study can be found by clicking on your group number. If you have forgotten your group number, you can find it in a file attached to the email containing instructions for this activity.

Section2_Q2 – Likert (4 items)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	Please rank your agreement to the following statements: <ol style="list-style-type: none"> 1. This assignment encouraged all group members to participate in the discussion (Section2_Q2A) 2. Students were not invited to share their ideas and knowledge by this assignment (Section2_Q2B) 3. Students were encouraged to ask questions to group members and were given meaningful answers in this assignment (Section2_Q2C) 4. Students were encouraged to express their own ideas to their group members in this assignment (Section2_Q2D)
---	--

Section2_Q3 – Open Ended

Question [open-ended]	In what ways did the content of this assignment (i.e. your group's assigned task and the data used) encourage collaboration among diverse group members and sharing of personal knowledge?
---------------------------------	---

Section2_Q4 – Open Ended

Question [open-ended]	In what ways did the content of this assignment discourage collaboration among diverse group members and sharing of personal knowledge?
---------------------------------	--

[Next Page]

During the lab activity, the group you were working in created a joint final answer. Please review this output to answer the next set of questions. Your final answer should be posted in your group's conversation log, which is located to the top right of the instruction guide in Udio.

Section2_Q5 – Likert (1 item)

Question [1-7 scale, 1 = strongly disagree, 7 = strongly agree]	Please rank your agreement with your group's final answers to the lab activity's question:
---	--

Section2_Q6 – Open ended

Question [Open-ended]	In approximately 200 words, please summarise what changes or additions you would make to your group's final answer:
---------------------------------	---

[Next Page]

SECTION 3 - EVALUATION OF INDIVIDUAL CONTRIBUTIONS

In this section, we would like you to consider your own participation and contributions to your group's discussion. Please keep in mind that the answers you provide in this section will not affect your participation grade for this activity. We welcome an honest reflection of your participation.

We ask that you review the chat record from the group activity at this time, which is located in Udio to the top right of the instruction guide.

First, we would like you to consider the **number or quantity** of messages you and your group members contributed to the conversation.

Section3_Q1 - Open Ended

Question [Short answer]	Approximately how many messages were contributed in total by the entire group throughout this activity?
-----------------------------------	--

Section3_Q2 – Open Ended

Question [Short answer]	Approximately how many messages did you contribute in total during this activity?
-----------------------------------	--

[Next Page]

We would now like you to consider the **quality** of your contributions to your group activity. Before answering the next questions, please refer to your group's conversation log and read through the contributions that you sent. This is located in Udio to the top right of the activity instructions.

Section3_Q3 – Open Ended

Question [Open ended]	How would you assess the quality of your own contributions to your group's discussion?
---------------------------------	--

Section3_Q4 – Open Ended

Question [Open-ended]	What factors (social, academic, etc) encouraged you to contribute to the group discussion?
---------------------------------	---

Section3_Q5 – Open Ended

Question [Open-ended]	What factors (social, academic, etc) discouraged you from contributing more to your group's discussion?
---------------------------------	--

Section3_Q6 - Likert (4 items)

Question [1-7 scale, 1 = Strongly disagree, 7 = Strongly agree]	<p>Please rank your agreement to the following statements, based on your own personal assessment of the quality of your contributions:</p> <ol style="list-style-type: none"> 1. My contributions during this activity were helpful to my group members. (Section3_Q6A) 2. I did not contribute my fair share of the assignment (Section3_Q6B) 3. My contributions were important to this group work activity (Section3_Q6C) 4. I participated in this activity at a level I believe is fair to my group members (Section3_Q6D)
---	---

[Next Page]

We now ask that you reflect on the entire group discussion. Before answering the next questions, please review your group's entire conversation log:

Section3_Q7 - Open Ended

Question	Based on the conversation log, what went well during your group's collaboration?
-----------------	--

[open-ended]	
--------------	--

Section3_Q8 – Open Ended

Question [open-ended]	Based on the conversation log, what were some difficulties or tensions that your group had (if any)?
---------------------------------	--

[Next Page]

SECTION 4 - EMOTIONAL REACTIONS TO DATA

For the next set of questions, we ask you to look at all the messages in the conversation log in Udio. When looking at all the messages, try to identify which message contains a **positive** or **negative** reaction to the data from the World Bank. Not all messages will easily fall into positive or negative reactions, however. Some messages may lack emotion and can be considered **neutral**. Some messages may contain both positive and negative content and should be considered **mixed**. Some messages may be difficult to determine if they are positive or negative, and these messages can be considered **ambiguous**. Here are some examples of sentences that are positive, negative, neutral, mixed, and ambiguous:

- **Positive:** “I really like that the Netherlands contributes 5.5% of their GDP for education”
- **Negative:** “The Netherlands only contributes 5.5% of their GDP for education which is not enough”
- **Neutral:** “The Netherlands contributes 5.5% of their GDP for education”
- **Mixed:** “It is good that the Netherlands contributes 5.5% of their GDP for education, but unfortunately they do not spend the money wisely”
- **Ambiguous:** “The Netherlands contributes 5.5% of their GDP for education, that seems pickles to me.”

Section4_Q1 – Open Ended

Question [integer]	Approximately how many POSITIVE messages did YOU contribute?
------------------------------	--

Section4_Q2 – Open Ended

Question [Integer]	Approximately how many POSITIVE messages did OTHERS contribute?
------------------------------	---

Section4_Q3 – Open Ended

Question [open-ended]	If you consider any messages to be positive please copy and paste between 1 and 3 messages that you believe are positive here.
---------------------------------	--

[Next Page]

Section4_Q4 – Open Ended

Question [Integer]	Approximately how many NEGATIVE messages did YOU contribute?
------------------------------	--

Section4_Q5 – Open Ended

Question [Integer]	Approximately how many NEGATIVE messages did OTHERS contribute?
------------------------------	---

Section4_Q6 – Open Ended

Question [open-ended]	If you consider any messages to be negative please copy and paste between 1 and 3 messages that you believe are negative here.
---------------------------------	--

[Next Page]

Section4_Q7 – Open Ended

Question [Integer]	Approximately how many MIXED messages did YOU contribute?
------------------------------	---

Section4_Q8 – Open Ended

Question [Integer]	Approximately how many MIXED messages did OTHERS contribute?
------------------------------	--

Section4_Q9 – Open Ended

Question [open-ended]	If you consider any messages to be mixed please copy and paste between 1 and 3 messages that you believe are mixed here.
---------------------------------	--

[Next Page]

Section4_Q10 – Open Ended

Question [Integer]	Approximately how many NEUTRAL messages did YOU contribute?
------------------------------	---

Section4_Q11 – Open Ended

Question [Integer]	Approximately how many NEUTRAL messages did OTHERS contribute?
------------------------------	--

Section4_Q12 – Open Ended

Question [open-ended]	If you consider any messages to be neutral please copy and paste between 1 and 3 messages that you believe are neutral here.
---------------------------------	--

[Next Page]

Section4_Q13 – Open Ended

Question [Integer]	Approximately how many AMBIGUOUS messages did YOU contribute?
------------------------------	---

Section4_Q14 – Open Ended

Question [Integer]	Approximately how many AMBIGUOUS messages did OTHERS contribute?
------------------------------	--

Section4_Q15 – Open Ended

Question [open-ended]	If you consider any messages to be ambiguous please copy and paste between 1 and 3 messages that you believe are ambiguous here.
---------------------------------	--

[Next Page]

Section4_Q16 – Open Ended

Question [Open-ended]	Please list the group members you believe in general had a POSITIVE reaction to the data from the World Bank.
---------------------------------	--

Section4_Q17 – Open Ended

Question [Open-ended]	Please list the group members you believe had in general a NEGATIVE reaction to the data from the World Bank.
---------------------------------	--

Section4_Q18 – Open Ended

Question [Open-ended]	Please list the group members you believe in general had a MIXED reaction to the data from the World Bank.
---------------------------------	---

Section4_Q19 – Open Ended

Question [Open-ended]	Please list the group members do you believe in general had a NEUTRAL reaction to the data from the World Bank information.
---------------------------------	--

Section4_Q20 – Open Ended

Question [Open-ended]	Please list the group members do you believe in general had an AMBIGUOUS reaction to the data from the World Bank information.
---------------------------------	---

Section4_Q21 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It was easy to categorize messages as positive, negative, neutral, mixed, and ambiguous.
---	--

Section4_Q22 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	Each group member likely categorized messages the same as me when judging them as positive, negative, neutral, mixed, and ambiguous. [1-7 scale: 1=strongly disagree, 7 = strongly agree]
---	--

Section4_Q23 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It is important that people do not misinterpret the emotional content of messages
---	---

Section4_Q24 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It would be helpful to know what category people selected when categorizing their own messages.
---	---

Section4_Q25 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It would be helpful to know what category people selected when categorizing my messages.
---	--

Section4_Q26 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It would be helpful if positive messages were highlighted in green in the discussion window.
---	--

Section4_Q27 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	It would be helpful if negative messages were highlighted in red in the discussion window.
---	--

Section4_Q28 Likert (1 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	After I have typed a message and before I sent it to the group. It would be helpful if a computer indicated to me privately a prediction about the emotional category of my message.
---	--

Section4_Q29 Open Ended

Question [Open-ended]	What do you think could be done to improve the clarity of emotional expression in group discussions?
---------------------------------	--

Section4_Q30 Open Ended

Question [Open-ended]	Do you think it is important to understand the emotional reactions of your group members during group work?
---------------------------------	---

[Next Page]

SECTION 5 - SOFT SKILLS AND SUPPORTS

We will now ask you to consider ‘soft skills’ of collaborative group work. Soft skills are essential non-academic skills that include areas such as: communication, decision making, motivation, leadership, teamwork, creativity or problem solving (among others). For the following questions, you may wish to refer to the articles in the ‘PBL Materials’ folder for Management of Organisations and Marketing (MOM) in your StudentPortal.

Section5_Q1 – Open Ended

Question	Which soft skills did you or your group members use when working with your small group for this assignment?
-----------------	---

[Open-ended]	
--------------	--

Section5_Q2 – Open Ended

Question [Open-ended]	Which particular soft skills are necessary for working with group members from diverse cultural backgrounds?
---------------------------------	--

Section5_Q3 – Open Ended

Question [Open-ended]	Which soft skills do you feel you should improve upon for your next group collaboration experience?
---------------------------------	---

[Next Page]

SECTION 6 - FEEDBACK TO THE MEMBERS OF MY LAB GROUP

One important goal of this assignment is for you to receive feedback from your group members that can help you strengthen your soft skills for collaborating with diverse groups of people. These are essential skills for those considering careers in business and economics fields.

We ask that you please provide polite, constructive feedback to each of your lab group members below, based on your group work experience and reflection of your group's conversation log. **This information will be shared with your group members. Your name will not be attached with the information shared.**

Providing quality, constructive feedback is an essential skill for business graduates. For guidelines about giving constructive feedback, you may wish to refer to resource listed below, which is available in the 'PBL Materials' folder for Management of Organisations and Marketing (MOM) in your Student Portal.

- Grohnert, T. (2015). *Giving and seeking constructive feedback and reflecting on your work.*

Section6_GroupMember1 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember1A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember1B)

<p>What actions or contributions did this group member make that particularly benefited the group in this assignment?</p> <p>[open-ended]</p> <p>(Section6_GroupMember1C)</p>
<p>What feedback would you provide to this group member for improving their contributions to group work activities in the future?</p> <p>[open-ended]</p> <p>(Section6_GroupMember1D)</p>

Section6_GroupMember2 – 3 Open Ended, 1 Likert

<p>Group member's username: (Section6_GroupMember2A)</p> <p>[short answer]</p>
<p>Please rate the helpfulness of this group member's contributions</p> <p>[1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful]</p> <p>(Section6_GroupMember2B)</p>
<p>What actions or contributions did this group member make that particularly benefited the group in this assignment?</p> <p>[open-ended]</p> <p>(Section6_GroupMember2C)</p>
<p>What feedback would you provide to this group member for improving their contributions to group work activities in the future?</p> <p>[open-ended]</p> <p>(Section6_GroupMember2D)</p>

Section6_GroupMember3 – 3 Open Ended, 1 Likert

<p>Group member's username: (Section6_GroupMember3A)</p> <p>[short answer]</p>
<p>Please rate the helpfulness of this group member's contributions</p> <p>[1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful]</p> <p>(Section6_GroupMember3B)</p>
<p>What actions or contributions did this group member make that particularly benefited the group in this assignment?</p> <p>[open-ended]</p> <p>(Section6_GroupMember3C)</p>
<p>What feedback would you provide to this group member for improving their contributions to group work activities in the future?</p> <p>[open-ended]</p> <p>(Section6_GroupMember3D)</p>

Section6_GroupMember4 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember4A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember4B)
What actions or contributions did this group member make that particularly benefited the group in this assignment? [open-ended] (Section6_GroupMember4C)
What feedback would you provide to this group member for improving their contributions to group work activities in the future? [open-ended] (Section6_GroupMember4D)

Section6_GroupMember5 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember5A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember5B)
What actions or contributions did this group member make that particularly benefited the group in this assignment? [open-ended] (Section6_GroupMember5C)
What feedback would you provide to this group member for improving their contributions to group work activities in the future? [open-ended] (Section6_GroupMember5D)

[tick box] I agree that my feedback provided above (in Section 6) can be shared anonymously with my individual group members

[Submit]

Thank you for completing this assignment! You have now finished all the required steps to receive your participation credit. You can expect to receive a feedback report on your own participation in this group work activity in approximately two weeks to your university email.

APPENDIX 5 - POST ACTIVITY FOR EXPERIMENT 2

This was adapted from the co-designed with Jenna Mittelmeier so some items are in support of her research interests while Section 4 represents my unique contribution in support of this PhD thesis work.

Thank you for participating in this post-activity.

You recently collaborated with a small group of your classmates in a computer lab on a case study related to education statistics. The following questions and activities is the second requirement for receiving your participation points for this week's assignment.

Throughout the activity, you will be asked to reflect upon the group work process and soft skills required to work with diverse group members. At the end of this session, you will be given the opportunity to provide constructive feedback about your group members' contributions to the activity.

In return, the feedback provided by your group members about your own participation in this assignment will be provided to your university email address in approximately two weeks. We hope that this feedback will serve as a useful tool for further developing essential soft skills and cross-cultural competencies for your future career.

Name

Question [short answer]	Name:
-----------------------------------	-------

StudentID

Question [short answer]	Student ID number:
-----------------------------------	--------------------

[Next Page]

SECTION 1 - REFLECTION AND RECALL OF GROUP WORK PROCESS

For the first section of this post-assignment, we would like you to reflect on the process of working together and the final outcome of your group assignment.

Section1_Q1 – Open Ended

Question [Open-ended]	In approximately 200 words, please reflect on the processes and outcomes of working together with your classmates during this computer lab assignment:
--	--

Section1_Q2 – Likert (4 items)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	<p>Please rank your agreement to the following statements:</p> <ol style="list-style-type: none"> 5. We have completed the task in a way we all agree upon. 6. I am not satisfied with the performance of our group. 7. I would wish to work with this group in the future 8. As a group, we have learned a lot.
---	--

Section1_Q3 – Likert (4 items)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	<p>Please rank your agreement to the following statements:</p> <ol style="list-style-type: none"> 1. The sentence frames were useful in supporting collaboration 2. The sentence frames were usable in supporting collaboration 3. The sentence frames were expressive in supporting collaboration 4. The sentence frames were effective in supporting collaboration
---	--

[Next Page]

SECTION 2 - REFLECTION OF ASSIGNMENT CONTENT

We would now like you to consider the content and topic of your assigned case study. You can view this again to refresh yourself by logging on to Udio at: <http://iet-projects.open.ac.uk/cethub>

Your Udio username is your first and last name (example: JohnSmith). The username is case sensitive.

Your password is your birthdate. For example, if you were born on July 1, 1996, your password would be 07011996.

Your case study can be found by clicking on your group number. If you have forgotten your group number, you can find it in a file attached to the email containing instructions for this activity.

Section2_Q2 – Likert (4 items)

<p>Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]</p>	<p>Please rank your agreement to the following statements:</p> <ol style="list-style-type: none"> 5. This assignment encouraged all group members to participate in the discussion (Section2_Q2A) 6. Students were not invited to share their ideas and knowledge by this assignment (Section2_Q2B) 7. Students were encouraged to ask questions to group members and were given meaningful answers in this assignment (Section2_Q2C) 8. Students were encouraged to express their own ideas to their group members in this assignment (Section2_Q2D)
---	---

[Next Page]

During the lab activity, the group you were working in created a joint final answer. Please review this output to answer the next set of questions. Your final answer should be posted in your group's conversation log, which is located to the top right of the instruction guide in Udio.

Section2_Q5 – Likert (1 item)

<p>Question [1-7 scale, 1 = strongly disagree, 7 = strongly agree]</p>	<p>Please rank your agreement with your group's final answers to the lab activity's question:</p>
---	---

[Next Page]

Section 3 - Evaluation of quality of discussion

In this section, we would like you to consider your own participation and contributions to your group's discussion. Please keep in mind that the answers you provide in this section will not affect your participation grade for this activity. We welcome an honest reflection of your participation.

We ask that you review the chat record from the group activity at this time, which is located in Udio to the top right of the instruction guide.

First, we would like you to consider the **number or quantity** of messages you and your group members contributed to the conversation.

[Next Page]

We would now like you to consider the **quality** of your contributions to your group activity. Before answering the next questions, please refer to your group's conversation log and read through the contributions that you sent. This is located in Udio to the top right of the activity instructions.

Section3_Q4 – Open Ended

Question [Open-ended]	What factors (social, academic, etc) encouraged you to contribute to the group discussion?
---------------------------------	---

Section3_Q5 – Open Ended

Question [Open-ended]	What factors (social, academic, etc) discouraged you from contributing more to your group's discussion?
---------------------------------	--

Section3_Q6 - Likert (15 items)

Question [1-7 scale, 1 = Strongly disagree, 7 = Strongly agree]	<p>What is the frequency with which you:</p> <ol style="list-style-type: none"> 1. communicated on your partner's emotions (F1) 2. adapted your behavior to your partner's emotions (F1) 3. communicated on your own emotions (F1) 4. understood your partner's emotions (F2) 5. imagined your partner's reactions to your emotions (F2) 6. compared your emotions to your partner's emotions (F2) 7. appeared able to control your own emotions (F2) 8. provided your own points of view (F3) 9. defended and argued your own ideas (F3) 10. built up on your partner's ideas (F3)
---	---

F1 = to communicate on emotions and adapt to emotions, F2 = to compare emotions and imagine reactions to emotion, F3 = to argue and build upon the other's ideas.

Section3_Q6 - Likert (15 items)

Question [1-7 scale, 1 = Strongly disagree, 7 = Strongly agree]	<p>What is the frequency with which your group members:</p> <ol style="list-style-type: none"> 1. communicated on your emotions (F1) 2. adapted their behavior to your emotions (F1) 3. understood their own emotions (F1) 4. communicated on their own emotions (F1) 5. adapted their behavior to their own emotions (F1) 6. imagined your reaction to their own emotions (F2) 7. compared their emotions to your emotions (F2)
---	---

	8. provided their own points of view (F3) 9. defended and argued their own ideas (F3) 10. understood your points of view (F3) 11. built up on your ideas (F3)
--	--

F1 = to communicate on emotions and adapt to emotions, F2 = to compare emotions and imagine reactions to emotion, F3 = to argue and build upon the other's ideas.

[Next Page]

We now ask that you reflect on the entire group discussion. Before answering the next questions, please review your group's entire conversation log:

Section3_Q7 - Open Ended

Question [open-ended]	Based on the conversation log, what went well during your group's collaboration?
---------------------------------	--

Section3_Q8 – Open Ended

Question [open-ended]	Based on the conversation log, what were some difficulties or tensions that your group had (if any)?
---------------------------------	--

[Next Page]

SECTION 4 - EMOTIONAL REACTIONS TO DATA

For the next set of questions, we ask you to look at all the messages in the conversation log in Udio. When looking at all the messages, try to identify which message contains a **positive** or **negative** reaction to the data from the World Bank. Not all messages will easily fall into positive or negative reactions, however. Some messages may lack emotion and can be considered **neutral**. Some messages may contain both positive and negative content and should be considered **mixed**. Here are some examples of sentences that are positive, negative, neutral, and mixed, and ambiguous:

- **Positive:** "I really like that the Netherlands contributes 5.5% of their GDP for education"
- **Negative:** "The Netherlands only contributes 5.5% of their GDP for education which is not enough"
- **Neutral:** "The Netherlands contributes 5.5% of their GDP for education"
- **Mixed:** "It is good that the Netherlands contributes 5.5% of their GDP for education, but unfortunately they do not spend the money wisely"

Section4_Q1 – Integer

Question [integer]	Approximately how many POSITIVE messages did YOU contribute?
------------------------------	--

Section4_Q2 – Integer

Question [Integer]	Approximately how many POSITIVE messages did OTHERS contribute?
------------------------------	---

Section4_Q3 – Open Ended

Question [open-ended]	If you consider any messages to be positive please copy and paste between 1 and 3 messages that you believe are positive here.
---------------------------------	--

[Next Page]

Section4_Q4 – Integer

Question [Integer]	Approximately how many NEGATIVE messages did YOU contribute?
------------------------------	--

Section4_Q5 – Integer

Question [Integer]	Approximately how many NEGATIVE messages did OTHERS contribute?
------------------------------	---

Section4_Q6 – Open Ended

Question [open-ended]	If you consider any messages to be negative please copy and paste between 1 and 3 messages that you believe are negative here.
---------------------------------	--

[Next Page]

Section4_Q7 – Integer

Question [Integer]	Approximately how many MIXED messages did YOU contribute?
------------------------------	---

Section4_Q8 – Integer

Question [Integer]	Approximately how many MIXED messages did OTHERS contribute?
------------------------------	--

Section4_Q9 – Open Ended

Question [open-ended]	If you consider any messages to be mixed please copy and paste between 1 and 3 messages that you believe are mixed here.
---------------------------------	--

[Next Page]

Section4_Q10 – Integer

Question [Integer]	Approximately how many NEUTRAL messages did YOU contribute?
------------------------------	---

Section4_Q11 – Integer

Question [Integer]	Approximately how many NEUTRAL messages did OTHERS contribute?
------------------------------	--

Section4_Q12 – Open Ended

Question [open-ended]	If you consider any messages to be neutral please copy and paste between 1 and 3 messages that you believe are neutral here.
---------------------------------	--

[Next Page]

Section4_Q15 – Likert (4 items

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	<p>Please rank your agreement to the following statements to describe your reaction to the main lab activity:</p> <ol style="list-style-type: none"> 1. I felt a mixture of emotions (Section4_Q15A) 2. I felt different emotions occur very quickly one after another (Section4_Q15B). 3. I felt only one thing throughout the event or experience*. (Section4_Q15C) 4. I felt either positive or negative emotions but not both at the same time*. (Section4_Q15D)
---	--

Section4_Q16 – Multiple Choice

Question [Open-ended]	<p>Would you describe your experience as</p> <ul style="list-style-type: none"> - POSITIVE - NEGATIVE
---------------------------------	---

	<ul style="list-style-type: none"> - MIXED - NEUTRAL
--	--

Section4_Q22 Likert (2 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	<ul style="list-style-type: none"> - It was easy to categorize messages as positive, negative, neutral, and mixed. - Each group member likely categorized messages the same as me when judging them as positive, negative, neutral, mixed, and ambiguous. - It is important that people do not misinterpret the emotional content of messages
---	--

Section4_Q23 Open Ended

Question [Open-ended]	What do you think could be done to improve the clarity of emotional expression in group discussions?
---------------------------------	--

Section4_Q24 Likert (2 item)

Question [1-7 scale: 1=strongly disagree, 7 = strongly agree]	<ul style="list-style-type: none"> - It would be helpful if positive messages were highlighted in green in the discussion window. - It would be helpful if negative messages were highlighted in red in the discussion window.
---	--

Section4_Q25 Open Ended

Question [Open-ended]	Do you think it is important to understand the emotional reactions of your group members during group work?
---------------------------------	---

[Next Page]

SECTION 5 - SOFT SKILLS AND SUPPORTS

We will now ask you to consider ‘soft skills’ of collaborative group work. Soft skills are essential non-academic skills that include areas such as: communication, decision making, motivation, leadership, teamwork, creativity or problem solving (among others). For the following questions, you may wish to refer to the articles in the ‘PBL Materials’ folder for Management of Organisations and Marketing (MOM) in your StudentPortal.

Section5_Q1 – Open Ended

Question [Open-ended]	Which soft skills did you or your group members use when working with your small group for this assignment?
---------------------------------	---

Section5 Q2 – Open Ended

Question [Open-ended]	Which soft skills do you feel you should improve upon for your next group collaboration experience?
---------------------------------	---

[Next Page]

SECTION 6 - FEEDBACK TO THE MEMBERS OF MY LAB GROUP

One important goal of this assignment is for you to receive feedback from your group members that can help you strengthen your soft skills for collaborating with diverse groups of people. These are essential skills for those considering careers in business and economics fields.

We ask that you please provide polite, constructive feedback to each of your lab group members below, based on your group work experience and reflection of your group's conversation log. **This information will be shared with your group members. Your name will not be attached with the information shared.**

Providing quality, constructive feedback is an essential skill for business graduates. For guidelines about giving constructive feedback, you may wish to refer to resource listed below, which is available in the 'PBL Materials' folder for Management of Organisations and Marketing (MOM) in your Student Portal.

- Grohnert, T. (2015). *Giving and seeking constructive feedback and reflecting on your work.*

Section6 GroupMember1 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember1A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember1B)
Would you describe this person's reaction to the world bank data as Positive, Negative, Neutral, Mixed, or Ambiguous?
What actions or contributions did this group member make that particularly benefited the group in this assignment? [open-ended]

(Section6_GroupMember1C)
What feedback would you provide to this group member for improving their contributions to group work activities in the future? [open-ended] (Section6_GroupMember1D)

Section6_GroupMember2 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember2A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember2B)
Would you describe this person's reaction to the world bank data as Positive, Negative, Neutral, Mixed, or Ambiguous?
What actions or contributions did this group member make that particularly benefited the group in this assignment? [open-ended] (Section6_GroupMember2C)
What feedback would you provide to this group member for improving their contributions to group work activities in the future? [open-ended] (Section6_GroupMember2D)

Section6_GroupMember3 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember3A) [short answer]
Please rate the helpfulness of this group member's contributions [1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful] (Section6_GroupMember3B)
Would you describe this person's reaction to the world bank data as Positive, Negative, Neutral, Mixed, or Ambiguous?
What actions or contributions did this group member make that particularly benefited the group in this assignment? [open-ended] (Section6_GroupMember3C)
What feedback would you provide to this group member for improving their contributions to group work activities in the future?

[open-ended]
(Section6_GroupMember3D)

Section6_GroupMember4 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember4A)
[short answer]

Please rate the helpfulness of this group member's contributions
[1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful]
(Section6_GroupMember4B)

Would you describe this person's reaction to the world bank data as Positive,
Negative, Neutral, Mixed, or Ambiguous?

What actions or contributions did this group member make that particularly benefited
the group in this assignment?
[open-ended]
(Section6_GroupMember4C)

What feedback would you provide to this group member for improving their
contributions to group work activities in the future?
[open-ended]
(Section6_GroupMember4D)

Section6_GroupMember5 – 3 Open Ended, 1 Likert

Group member's username: (Section6_GroupMember5A)
[short answer]

Please rate the helpfulness of this group member's contributions
[1-7 scale, 1 = extremely unhelpful, 7 = extremely helpful]
(Section6_GroupMember5B)

Would you describe this person's reaction to the world bank data as Positive,
Negative, Neutral, Mixed, or Ambiguous?

What actions or contributions did this group member make that particularly benefited
the group in this assignment?
[open-ended]
(Section6_GroupMember5C)

What feedback would you provide to this group member for improving their
contributions to group work activities in the future?
[open-ended]
(Section6_GroupMember5D)

[tick box] I agree that my feedback provided above (in Section 6) can be shared anonymously with my individual group members

[Submit]

Thank you for completing this assignment! You have now finished all the required steps to receive your participation credit. You can expect to receive a feedback report on your own participation in this group work activity in approximately two weeks to your university email.

APPENDIX 6 – HREC FOR 2016

HUMAN RESEARCH ETHICS COMMITTEE (HREC) PROFORMA

Open University research involving human participants or materials has to be reviewed and where necessary, agreed by the HREC. To apply to HREC, please complete and email this proforma to Research-REC-review@open.ac.uk. You will need to attach any related documents for example: a consent form, information sheet, a questionnaire, consent form, or publicity leaflet, so that the HREC Review Panel has a full application. Ensure that if you have more than one group of participants, that the relevant documents for each research group are included. Omitting any documents may result in a delay to the review and approval process. No potential participants should be approached to take part in any research until you have received a response from the HREC Chair.

If you have any queries about completing the proforma please look at the Research Ethics website, in particular the FAQs - <http://www.open.ac.uk/research/ethics/faq-questions-inline> which includes sample documents and templates. You can also contact the [HREC Chair](#) or [Secretary](#).

The submission deadline for applications is **every Thursday at 5.30pm** when they will be assessed for completeness and then sent to the HREC Review Panel. Once an application has been passed for review you should receive a response within 21 working days.

All general research ethics queries should be sent to Research-Ethics@open.ac.uk.

PLEASE COMPLETE ALL THE SECTIONS BELOW – DELETING THE INSERTED INSTRUCTIONS

Project identification and rationale

1. TITLE OF PROJECT

2. ABSTRACT

Recent findings in neuroscience suggest that we should be examining the intersection of emotion and cognition (Okon-Singer et al., 2015) and the implications for education ask us to consider that learning simply might be an emotional experience (Immordino-Yang, 2016; Immordino-Yang & Damasio, 2007). While this has broad implications for educational research in the context of online learning, several researchers are seeking to explore increasing emotional awareness in online learning environments (Arguedas et al., 2016; Feidakis et al., 2014). When exploring supporting written expression in online communication there are a variety of strategies which include prompts such as sentence starters that support communication by providing a stem to support writing (Morris et al., 2010). The goal of this RCT study is to examine an intervention of using emotional (n= 120) versus cognitive (n = 120) framing sentence starters on online communication to enable clarifying the emotional aspects of a written statement. The sentence stems will encourage people to lead their statements with an emotional categorization preface (i.e. "I had a [emotional category] reaction because..."). This could potentially improve communication by reducing the amount of ambiguity in a discussion – one of the limits to our interpretation of the emotional aspects of written expression. As the study is designed to provide a valuable learning experience with opportunities for reflection and feedback, with participants having options to opt out before, during, and after the study, there are limited risks to participants.

Project personnel and collaborators
--

3. INVESTIGATORS

Give names and institutional attachments of all persons involved in the collection and handling of individual data and name one person as Principal Investigator (PI). Research students should name themselves as PI and it is a requirement that a brief separate supervisor endorsement is sent to Research-REC-Review@open.ac.uk to support the application. This needs to be received with the application or shortly after, as the application cannot be processed without it (see

[Applications from research students: supervisor endorsement](#)). Please include the relevant HREC reference number in the subject line.

Principal Investigator/
(or Research Student):

Garron Hillaire

Other researcher(s):

Jenna Mittelmeier

Bart Rienties

Dirk Tempelaar

Bart Rienties

Primary Supervisor (if
applicable):

Research protocol

4. LITERATURE REVIEW

Context

University students are often early adopters of new socially networked communication technologies (Choi, 2016; Gallardo-echenique, Bullen, & Luis Marqués-Molíás, 2016). One popular social networking site noted recently that people had a need to express emotional reaction that were more complex than just liking something by introducing emotional reactions with icons like, love, haha, wow, sad, angry (Stinson, 2016). This functionality in Facebook is not very dissimilar to selecting from a list of emotional language for self-report (Ruiz, Charleer, Fernández-castro, & Duval, 2016). When interviewing students about emotional awareness in online group collaboration a qualitative study concluded that for emotion communication “[s]tudents need to provide other group members with a much more detailed textual description in order to fully simulate the communicative richness of a face-to-face encounter” (Robinson, 2013).

What is known

A recent study at the OU illustrates that when examining student satisfaction and learning design online communication with peers and tutors is strongly correlated with course completion (Rienties & Toetenel, 2016). When taking the context of emotion expression into consideration, the emotional aspect of communication merits further exploration as a part of investigation into the role of emotions in learning and student engagement. One common learning strategy to support guided practice of written expression is the use of sentence starters (Hall & Vue, 2012). Sentence starters (aka sentences openers) are considered a specific type of prompt (Weinberger et al., 2005). The literature on the efficacy of prompts is inconclusive in that some findings are positive while others have mixed effects (Morris et al., 2010). A potential factor on whether or not a prompt produces better collaboration is the balance between structure and flexibility (Ge & Land, 2004; Morris et al., 2010). This is because too much structure may in fact impede collaboration (Morris et al., 2010). There are considered to be three categories of prompts: procedural, reflection, and elaboration (Ge & Land, 2004). Sentence starters have been used in emotional journaling to help identify learner dispositions

Culture and Social Considerations

It is well known that there are cultural differences in determining the appropriateness of emotional communication (Bagozzi, Wong, & Yi, 1999; Shao, Doucet, & Caruso, 2014). The frame of valence has been examined in cross cultural settings and found that the dimension of valence appear to be universal in the sense that people across cultures can identify emotions as either positive or negative (Russell, 1983, 1991). There are also known display rules that we in turn use as social cues for what emotion is appropriate in the current context (Malatesta & Haviland, 1982).

Emotions in Learning

The idea of contextualizing emotions in education has led to a variety of strategies that focus on examining discrete emotions that are relevant to learning (D'Mello et al., 2014; Pardos, Baker, & Pedro, 2013; Peterson, Brown, & Jun, 2015). For example, Pekrun, Frenzel, & Goetz (2007) outlined that for education the most relevant emotions are the set of discrete emotions, called achievement emotions, which are tied directly to achievement activities or achievement outcomes. However, proponents for core affect outline that the dimensional approach is not limiting in terms of narrowing the potential emotional communication to discrete emotions (Barrett, 2006). It is important to note that while literature can, at times, be organized in terms of positive and negative emotions, it is believed that contrasting emotions in learning is favored over a hedonistic perspective (Fiedler & Beier, 2014). However, there is utility in thinking about the positive and negative emotions in the context of Piaget's theory of constructivism as positive emotions align well with the concept of assimilation while negative emotions align well with accommodation (Fiedler & Beier, 2014). Dialectical emotions have been shown to have a correlation with critical thinking (Hui et al., 2009). One potential explanation for this is that in dialectical experiences students could potentially benefit from emotional support to both assimilate and accommodate information.

Self and Socially Regulated Learning

In online learning it has been pointed out that there is typically more autonomy for the students making it an interesting context to study self-regulated learning (Rienties et al., 2012). However, some have made the observation (Järvelä, 2014) that self-regulated learning was established in a context that has more of a focus on the individual differences (Zimmerman & Martinez-pons, 1990) while online collaborative environments include more social regulation sparking an extension of SRL to frame research in terms of Self and Socially Regulated Learning (SSRL) (Järvelä, 2014). Emotions have been considered to be essential for self-regulated learning (Op ' et al., 2007) and at the same time have been considered something that are socially regulated (Reeck et al., 2016). Emotions have also been acknowledged as an underexplored facet of SSRL (Järvelä, 2014; Järvelä et al., 2013).

Gaps in the Research

One controversial universal theory on emotions is core affect, which categorized emotions on the dimensions of valence, arousal, and sometimes control (Russell, 1980; Russell & Barrett, 1999b; Yik, Russell, & Steiger, n.d.). While it is a controversial theory it is very influential as the dimensions are frequently used to organize self-report of emotional response (Dan-Glauser & Scherer, 2011; Västfjäll & Gärling, 2007). What is not understood is if this categorization strategy could potentially provide an appropriate balance of structure and flexibility for sentence starters. The goal of this study is to examine an intervention of using emotional versus cognitive framing sentence starters on online communication to enable clarifying the emotional aspect of a written statement. The sentence stems will encourage people to lead their statements with an

emotional categorization preface (i.e. “I had a [emotional category] reaction because...”). The categories used in the sentence stems will be: Positive, Negative, Neutral, and Mixed (e.g. “I had a positive reaction because...”). This could potentially improve communication by reducing the amount of ambiguity in a discussion – one of the limits to our interpretation of the emotional aspect of written expression (Barchard, Hensley, Anderson, & Walker, 2013; Wilson et al., 2005).

5. METHODOLOGY

This study is part of a joined research project with Jenna Mittelmeier (HREC/2360). This study will take place in a computer lab setting at Maastricht University, where participants will work online in groups of five on one of three collaborative tasks (which are attached to this application). The tasks will require participants to evaluate and discuss the open education data set provided online by the World Bank EdStats Dashboard (<http://datatopics.worldbank.org/education/>). The collaboration will occur within the Udio software system (described below), with participants working exclusively within an asynchronous service. This research project will incorporate a randomised control trial method that supports two conditions and a control. The control condition will not intervene in the group work element of the task and will follow the design of Jenna Mittelmeier. Condition 1 will intervene in the group work task asking participants to reflect on something important to the learning goal and communicate the important point to their peers with the sentence stem “An important point to consider is [important point] because...” in the second condition the emotional sentence frame will be used “I had a [category of emotion] reaction to [important point] because...”. The two conditions will help to determine if asking participants to communicate what they find important as well as examining if framing the important point with an emotional response might improve collaboration by providing additional insight into why the participant found the point to be important.

N=120	N=120	N=120
Control Group	Condition 1	Condition 2

Given that the communication strategy is to use a categorization that is considered to be universal, an attempt will be made when creating groups of 5 participants to have the composition be cross-cultural. The 360 participants in these conditions will be a part of a larger study where the content is varied between conditions creating a context of 1200 overall participants.

In the control group participants will work on the group tasks as previously described by Jenna Mittelmeier (HREC/2360).

In condition 1, participants will be asked to use the following prompt 2 times during the group discussion:

“An important point to consider is ...”

In condition 2, participants will be asked use of the following prompts 2 times during the group discussion

- I had a negative reaction to ...
- I had a positive reaction to ...
- I had a mixed reaction to ...
- I had a neutral reaction to ...

Participants will also be debriefed at the end of the lab (explained in detail later in this application). A rough outline of the individual lab timelines in provided below. Tentative lab schedule for each hour (70 minutes)

5 minutes Getting settled and informed consent
5 minutes Pre-test Col survey
5 minutes Explanation of task
40 minutes Time for group collaboration in Udio
5 minutes Post-test Col survey
4 minutes Wrap up and debrief

24 hours after the lab exercise there will be a 2 hour post lab activity that will ask participants to recall their experiences with the cognitive group outcomes of the lab activity. This post lab activity is part of the learning activity for that particular week and students are expected to participate both in the lab and post-lab activity, unless they opt for the alternative assignment (see section 7). Afterwards, they are asked to critically analyse the group output and make recommendations for further refinements of the learning outcomes. These cognitive outcomes will be shared with respective group members and provide valuable learning opportunities. As a next step, participants will be asked to recall their emotional experience, examine the emotional aspect of written expression from the group exercise, and examine the emotional aspect of the work product produced by the group. In the post activity participants will be asked to recall the lab work and be asked to label their interactions with the learning content as either positive, negative, neutral, or mixed. After making that judgement they will then be asked to log into Udio and review the chat log from the group activity. After reviewing the chat log, participants will be asked to estimate how many comments were made by the entire group together. Then estimate how many comments they personally contributed to the discussion. Then estimate how many comments they made were positive, how many were negative, how many were neutral, and how many were mixed. After providing an estimate that is greater than zero each participant will be asked to provide what they believe the be the best example(s) by copy and pasting one or two comments they made that have this emotional aspect. After analyzing the emotional aspect of their comments participants will be asked if they think their peers perceived them to have a positive, negative, neutral, or mixed response to the learning material. Finally, participants will be asked to review what was produced in the learning activity. After reviewing what the group submitted the participant will be asked if what was produced was positive, negative, neutral, or mixed. Also they will be asked to rate the outcome of their work on a scale from 1 to 7 whether or not the outcome represented critical thinking by weighing multiple evidence sources, making judgements about the evidence, and generating a defensible opinion.

The methodology will be quantitative in nature exploring the relationship between emotion and cognition. The basic overarching methodology will be feature generation for text classification (Forman, 2007) which can be fed into statistical models to answer questions about the relationship between the production of categorized emotional expression in text within the context of individuals and groups as it relates to outcome measures (see Appendix for description of proposed Natural Language Processing techniques).

The emotional aspect of written expression will be compared to the outcome variables and other trace data (e.g. self-report) to help identify if any specific category of emotion expression is correlated with improved learning outcomes. While the expression will be traced at the individual comment level, a hierarchical linear model (HLM) will be used as the theoretical framework of investigation is self and socially regulated learning. HLM was introduced into education research to examine the effects that group contexts have on students (Burstein, 1980). Observation and video capture (where applicable) of participant will be used to provide an additional predictor of emotional response of participants. The video data will be post processed by affective computing techniques to predict emotional response. It has been noted that including some measure of human observation is good practice in emotions studies (Mauss & Robinson, 2009). While human observers will be in the classroom video of a subset of participants will be collected to generate another trace of data that can benefit from human observation. Similarly the text generated from discussion will be coded by human reviewers to get another form of human observation incorporated into the sentiment analysis (Wilson et al., 2005) which can be used as a means of verifying that sentiment analysis is not too far from machine coding.

For the HLM the data will be categorized as background process and outcome (Burstein, 1980). Background data will include demographic information, prior student performance data (e.g. previous grades in the course). The process data will be the text features generated from NLP that categorizes comments on the core affect dimensions of emotion. The outcome data will be based on evaluations of the assignment from this task administered later in the course. The self-evaluation of critical thinking will be used in conjunction with an evaluation of the assignment conducted by researchers scoring the assignment on the presence of critical thinking.

6. PARTICIPANTS

Participants are first year university students at Maastricht University, who are taking a course with Dr Dirk Tempelaar. This study builds on previous work of Simon Knight (HREC/2014/66836/Knight/2) and Jenna Mittelmeier (HREC/2360). Participants in his module will be sent an email by their Dr Tempelaar to participate in a computer laboratory study session, which will be built into the course schedule (see next section for details). This classroom is highly diverse, with international students present from countries around Europe and the world. Thus, it is expected that participants will be come from a wide range of backgrounds, including domestic Dutch students, European students, and students from outside of Europe.

Altogether around 1200 participants are expected. All procedures and materials for this study were negotiated with our collaborator and gatekeeper at Maastricht University (Dr Dirk Tempelaar).

7. RECRUITMENT PROCEDURES

This study is built into the schedule of an introductory statistics class. Participants can opt to take part in the study, or to complete a separate course-related task of equal time and effort. Thus, participants are not required to attend the study or participate in the lab session. They may also withdraw and complete the alternative task at any time. It is worth noting that participants will not receive grades or marks for study participation or the alternative task (i.e. they will not be penalized for minimal effort or completion). Therefore, the study will not directly impact participant's grade or courses. Potential participants will receive an email from their classroom teacher one week prior to the lab, which will include our study information sheet (attached). They will be made aware at this time of the alternative assignment, as well as the notion that they are not required to attend and will not be marked for this activity.

8. CONSENT

Participants in this study will be prompted to give informed consent at the start of the lab activity. They will first be provided with a verbal briefing upon arrival, accompanied by a full online consent form (attached to this application). Participants will give consent by ticking an on-screen box, logging into the Udio system (using a unique log-in populated for individual participants), and submitting their unique ID number. In doing so, the participants will also consent to Maastricht University sharing demographic data about them with the Open University research team. Participants will be unable to progress into the activity without providing informed consent.

9. LOCATION(S) OF DATA COLLECTION

This study will take place at the Maastricht University in Maastricht, Netherlands, with participants in a computer lab using the Udio software system (explained below). Servers for the software system will be hosted securely by the Institute of Educational Technology on a Postgres system within the UK, however, meaning data collection from the server will take place on the OU campus. No Udio data will be stored on any servers or computers at Maastricht University as browser cache will be cleared after each lab session. Demographic data about participants who

agree to take part in this study will be collected from the host university (including age, gender, nation of origin, nation of citizenship). This will be sent from the classroom teacher (Dr Dirk Tempelaar) by an encrypted, password protected file via email. Participant identities in this file will be anonymised through the use of unique identification numbers in line with the Data Protection Act's definition of 'anonymized.'

Udio Description

Udio is a literacy platform developed by CAST.org, funded by the Office of Special Education Programs from the U.S. Department of Education (<http://cet.cast.org/udio/>). The platform was intended to create a curated core set of high interest reading material supplied by content partners in order to provide literacy supports for all students, including students with disabilities. While the platform was originally developed with the middle school population in mind, the Udio platform itself has a great potential to be used as a more general research platform. In partnership with CAST, researchers at the OU will explore the potential benefits of Udio as a research platform in a higher education context. The Udio platform functions within a web browser, and it is not a standalone software. Any material can be provided on the Udio platform as long as it can be formatted into an XML structure, making it possible to take most content that was developed for the internet and create a version of that material within Udio. When content is reformatted into Udio, it is referred to as an 'article.' Articles can be either single page or have formatting like chapter books with a table of contents. For a single page article there are supports on the page. The set of support features include:

- * a discussion feature that is displayed side by side with the content as well
- * a reading comprehension check
- * the ability to collection snippets of text or images from articles in Udio
- * the ability to create and publish a project based on material in Udio

For the purposes of this study, the assignments for the class will be provided as a single page article in Udio with all features available on the page of the assignment. The supports we will focus on using are the discussion supports as well as the project feature. During the lab Upon arrival to the lab, participants will be given an introduction to the activity and be prompted to provide informed consent (further described in a later section). Participants will then be asked to log into the Udio system on a web browser using a unique log in. Inside the system, they will be asked to read their activity instructions and begin the group work activity using the discussion feature. The Udio platform will be hosted on OU servers and this installation will be used for this study.

10. SCHEDULE

This study will take place October 10 – 14, 2016 at Maastricht University. The data will be analysed and written up for the primary investigator's PhD as part of study 1 by April 2017.

Key Ethics considerations

11. PUBLISHED ETHICS AND LEGAL GUIDELINES TO BE FOLLOWED

BERA

12. DATA PROTECTION AND INFORMATION SECURITY

This study has been registered with the University's Data Protection Coordinator. All data related to this study will be saved in a password-protected file on a university computer. All personally identifying information (such as name and Participant ID) will be removed, but unique identifier numbers will be used in line with the Data Protection Act's definition of 'anonymized.' Destruction of the data will occur at the earliest in October 2018 (or the end of this PhD project) and at the latest by October 2020, in order to allow for journal article publication and dissemination of findings.

13. RESEARCH DATA MANAGEMENT

The Udio Platform will collect log data stored on a Postgres database. During the study, all clickstream data will be logged around what materials were accessed, as well as traces of support utilization including discussion comments. At the conclusion of the study there will be a timeframe of 4 weeks where the data will remain on the server, allowing for data pre-processing using the database to format and extract data into a format ideal for conducting statistical analysis. During this month, some locally-hosted tools will be used to access data from the database for exploratory analysis purposes. Once that data have been prepared for statistical analysis, a backup of the database will be created for archival purposes to ensure that the research is reproducible. This archive will be stored in a secure manner on OU servers and maintained through 2020. Once archived, the database will only be used in the event that additional information is needed for retrieval to conduct a secondary analysis. Each of the investigators on the project will pull a pre-processed set of data from the database prior to archiving the database. The data will be to explicitly support the research questions outlined in the studies. Each of the researchers will take their data files and manage them appropriately in terms of keeping the files on OU machines and following OU data security policies. These independent files will for analysis purposes will be retained by research staff through 2020.

14. DECEPTION

None

15. RISK OF HARM

None

16. DEBRIEFING

A debriefing will be given verbally to participants at the end of the lab activity. In addition, students will receive cognitive feedback from their peers on the group outcomes and feedback from the post-lab activities, which will be a useful learning experience. A summary of the research findings will be compiled at the end of the analysis phase of this study and a copy of this report will be shared with participants via email, as well as with the Business & Economics department at Maastricht University. This report will contain no identifying information about participants. Participants will also be informed that they may contact the PI with any questions after the lab, or if they would like to withdraw part or all of their data up to 90 days afterwards. They will be given a copy of the study information sheet for their personal records, which includes contact details of the research team.

Project Management

17. RESEARCH ORGANISATION AND FUNDING

This study is part of the Leverhulme Open World Learning PhD programme.

18. OTHER PROJECT-RELATED RISKS

None

19. BENEFITS AND KNOWLEDGE TRANSFER

This research aims at better understanding of how to support emotional content to communication in online collaboration as well as how that relates to learning measurements and outcomes. There are practitioner implications as the sentence stem intervention is based on effective teacher practice methods. There are analytics implications as the use of sentiment analysis to detect emotional aspect of written communication will be examined. Finally there are research implications as the role of emotion expression in written communication will be examined in a group work context.

20. DISSEMINATING AND PUBLISHING RESEARCH OUTCOMES

The findings of this study will also be compiled for publication in the Computers & Education and the Journal of Computer Assisted Learning. Conference articles will also be written for EARLI, CSCL, and LAK. All publications will be available on ORO and will be circulated via social media.

21. DECLARATION

I declare that the research will conform to the above protocol and that any significant changes or new ethics issues will be raised with the HREC before they are implemented.

I declare that I have read and will adhere to the following two OU documents:

- [OU Code Of Practice For Research and at the Open University](#)
- [OU Ethics Principles for Research involving Human Participants](#)

<http://www.open.ac.uk/research/ethics/index.shtml>)

To meet internal governance and highlight OU research, the titles of all projects considered by the HREC (whether by HREC checklist or proforma), will be added to the Research Ethics website - <http://www.open.ac.uk/research/ethics/human-research>. If you would prefer for your title **not** to be made public, or have any queries, please email the HREC Secretary on Research-REC-Review@open.ac.uk.

	Garron Hillaire
Name:	_____
	IET
Unit/Faculty:	_____
	07493079493
Telephone	_____
	Garron.hillaire@open.ac.uk
E-mail	_____
	Garron Hillaire
Signature(s) (this can be the typed name(s) of investigator(s) if an electronic copy is submitted (which is preferred))	_____
	August 5 th , 2016
Date:	_____

End of project final report

Once your research has been completed you will need to complete and submit a final report to the HREC. A copy of the template can be found on the Research Ethics website at <http://www.open.ac.uk/research/ethics/human-research/human-research-ethics-full-review-process-and-proforma#final-report>.

May 2018

Proposed date for final report: _____

APPENDIX

In order to generate features about text Natural language processing (NLP) is the application of computational methods to process written expression which includes the sentiment analysis or opinion mining. Sentiment analysis attempts to identify written expresses as positive or negative (Pang & Lee, 2006). Sentiment has a distinction from emotion in that emotions can be free floating while sentiment typically has a target object (Munezero, Montero, Sutinen, & Pajunen, 2014).

Sentiment analysis has been used to categorize the emotional aspects of written expression, sentiment, mood, and at times model discrete emotions (Roan et al., 2009). The approach has a variety of methods including lexical approaches and machine learning approaches (Medhat, Hassan, & Korashy, 2014). Lexical approaches are considered to be the simplest approach. In lexical approaches dimensions like valence (positive to negative) are computing using word substitution to score sentences by averaging the score of the words as they are previously ranked in a dictionary on the dimension of valence (Pang & Lee, 2006). Dialectical emotional complexity will be identified by comparing collocates of words with opposing values based on the dictionary substitution method (Grossmann, Huynh, & Ellsworth, 2015). The approach can become more complex with methods that use machine learning classifiers (Medhat et al., 2014).

APPENDIX 7 – PERMISSION FORM FOR 2016

Thank you for participating in this lab assignment.

Activity data for this assignment will be collected for research purposes. Below is information about this joint study.

Study title: Investigating tools to support emotion communication during cross-cultural collaborative group work

What is the purpose of the study?

We are inviting you to take part in a study evaluating how diverse groups work together and if different types of academic content play a role in task behaviour and discourse as well as the potential for sentence starters to clarify communication.

Why have I been approached?

For the purposes of the study we need to recruit a number of groups of students studying in a higher education institution.

Do I have to take part?

No. Participation is entirely voluntary. If you change your mind about taking part in the study you can withdraw at any point during this session and at any point up to 90 days after the session. If you decide to withdraw, all your data will be destroyed and will not be used in the study. There are no consequences to deciding that you no longer wish to participate in this laboratory activity. There is an alternative assignment from your teacher for those who do not wish to participate.

What happens during the study?

You will be able to complete this study from a quiet location in a computer lab where you won't be disturbed. You will be working in small groups of five participants to complete a problem based learning task. You will be working in a different location to your group members. Altogether, this lab activity will take 70 minutes, and you will be asked to complete a post-activity at home which will take approximately two hours.

The study will involve collaboration via an online instant messaging system in your web browser with group members for no longer than 45 minutes. During this time, you will be asked to review and reflect on educational data available from web resources. During the post-activity task, you will be asked to reflect on your group's collaboration process and soft skills of working with diverse peers. Some participants will be selected to use sentence starters that support emotional and/or cognitive elements of communication.

To complete this task, you will be asked to log in a website called Udio. While you are logged in, the website will collect data of your instant messaging conversation. In collaboration with Dr Dirk Tempelaar we will collate anonymised demographic for analysis about how students learn together, including your gender, age and nation of origin. Some participants will be video recorded. The reason for video capture is to look for emotional reactions during discussions by observing the video, using artificial intelligence that interprets facial expression, and artificial intelligence that predicts heart rates based on techniques like motion capture. If you are located at a desk where a web camera is recording you will be notified at the start of the lab and asked if you would like the cameras turned away during the activity to avoid being recorded.

What are the possible disadvantages and risks of taking part?

We do not anticipate any risks associated with participation in this study.

What are the possible benefits of taking part?

You will gain an insight into how a psychology research project is conducted and what it is like to be a participant in such a study. The tasks are relevant to all students as they are about the kind of transferable skills around collaboration, inferring about data and finding information, that all graduates should have. You will also receive individualised feedback about your contributions to your group, including relevant soft skills for collaboration.

Will my taking part in this study be kept confidential?

Yes, no personally identifying information will be shared.

What will happen to the results of the research study?

You will be given a personal login for the Udio system. Identifying information will be kept passworded, and will not be associated with the anonymous data collected.

This research forms part of Jenna Mittelmeier's doctoral research at the Open University supervised by Dr Bart Rienties and Prof Denise Whitelock. It also forms part of Garron Hillaire's doctoral research at the

Open University supervised by Dr. Bart Rienties, Prof. Mark Fenton-O'Creevy, and Prof. Zdenek Zdrahal. All data will be available to Dr Dirk Tempelaar in his capacity as course leader.

Deanonimised data will not be shared other than within the OU supervisory and external examiner team, except where we are legally bound to do so. Pseudonyms will be used in reporting, and any identifying information mentioned in the instant messenger logs will be redacted.

The data will be kept in full for the duration of the investigator's PhD research or until December 2018 (whichever is later). Data stored will be kept in a password protected file in accordance with the Data Protection Act.

Who is organising and funding the research?

The research is organised by Jenna Mittelmeier & Garron Hillaire, who are research students at the Open University's Institute of Educational Technology. This work is funded by the Open University and the Leverhulme Trust.

Who has reviewed the study?

The Open University Ethics Committee has reviewed and approved this study.

Contact for Further Information

Jenna Mittelmeier

Institute of Educational Technology, Walton Hall, Milton Keynes, MK7 6AA

Email: jenna.mittelmeier@open.ac.uk

Garron Hillaire

Institute of Educational Technology, Walton Hall, Milton Keynes, MK7 6AA

Email: Garron.hillaire@open.ac.uk

Agreement to participate

[Tick box here]

I understand that checking this box constitutes a legal signature confirming that I acknowledge and agree to the above terms.

[Finish button]

APPENDIX 8 – HREC FOR 2017

HUMAN RESEARCH ETHICS COMMITTEE (HREC) PROFORMA

Open University research involving human participants or materials has to be reviewed and where necessary, agreed by the HREC. To apply to HREC, please complete and email this proforma to Research-REC-review@open.ac.uk. You will need to attach any related documents for example: a consent form, information sheet, a questionnaire, consent form, or publicity leaflet, so that the HREC Review Panel has a full application. Ensure that if you have more than one group of participants, that the relevant documents for each research group are included. Omitting any documents may result in a delay to the review and approval process. No potential participants should be

approached to take part in any research until you have received a response from the HREC Chair.

If you have any queries about completing the proforma please look at the Research Ethics website, in particular the FAQs - <http://www.open.ac.uk/research/ethics/faq-questions-inline> which includes sample documents and templates. You can also contact the [HREC Chair](#) or [Secretary](#).

The submission deadline for applications is **every Thursday at 5.30pm** when they will be assessed for completeness and then sent to the HREC Review Panel. Once an application has been passed for review you should receive a response within 21 working days.

All general research ethics queries should be sent to Research-Ethics@open.ac.uk.

PLEASE COMPLETE ALL THE SECTIONS BELOW – DELETING THE INSERTED INSTRUCTIONS

Project identification and rationale

1. TITLE OF PROJECT

Replication of Maastricht RCT of Emotion Sentence Starters by Garron Hillaire (Previous approval HREC/2388).

2. ABSTRACT

Recent findings in neuroscience suggest that we should be examining the intersection of emotion and cognition (Okon-Singer et al., 2015) and the implications for education ask us to consider that learning simply might be an emotional experience (Immordino-Yang, 2016; Immordino-Yang & Damasio, 2007). While this has broad implications for educational research in the context of online learning, several researchers are seeking to explore increasing emotional awareness in online learning environments (Arguedas et al., 2016; Feidakis et al., 2014). When exploring supporting written expression in online communication there are a variety of strategies which include prompts such as sentence starters that support communication by providing a stem to support writing (Morris et al., 2010). The goal of this RCT study is to replicate the pilot study in order to validate a sentiment analysis measure. The sentence stems will encourage people to lead their statements with an emotional categorization preface (i.e. “I had a [emotional category] reaction because...”). This could potentially improve communication by

reducing the amount of ambiguity in a discussion – one of the limits to our interpretation of the emotional aspects of written expression. As the study is designed to provide a valuable learning experience with opportunities for reflection and feedback, with participants having options to opt out before, during, and after the study, there are limited risks to participants.

Project personnel and collaborators

3. INVESTIGATORS

Give names and institutional attachments of all persons involved in the collection and handling of individual data and name one person as Principal Investigator (PI). Research students should name themselves as PI and it is a requirement that a brief separate supervisor endorsement is sent to Research-REC-Review@open.ac.uk to support the application. This needs to be received with the application or shortly after, as the application cannot be processed without it (see [Applications from research students: supervisor endorsement](#)). Please include the relevant HREC reference number in the subject line.

Principal Investigator/
(or Research Student):

Garron Hillaire

Bart Rienties

Dirk Tempelaar

Other researcher(s):

YingFei Heliot

Bart Rienties

Primary Supervisor (if
applicable):

Research protocol

4. LITERATURE REVIEW

Context

University students are often early adopters of new socially networked communication technologies (Choi, 2016; Gallardo-echenique et al., 2016). One popular social networking site noted recently that people had a need to express emotional reaction that were more complex than just liking something by introducing emotional reactions with icons like, love, haha, wow, sad, angry (Stinson, 2016). This functionality in

Facebook is not very dissimilar to selecting from a list of emotional language for self-report (Ruiz et al., 2016). When interviewing students about emotional awareness in online group collaboration a qualitative study concluded that for emotion communication “[s]tudents need to provide other group members with a much more detailed textual description in order to fully simulate the communicative richness of a face-to-face encounter” (Robinson, 2013).

What is known

A recent study at the OU illustrates that when examining student satisfaction and learning design online communication with peers and tutors is strongly correlated with course completion (Rienties & Toetenel, 2016). When taking the context of emotion expression into consideration, the emotional aspect of communication merits further exploration as a part of investigation into the role of emotions in learning and student engagement. One common learning strategy to support guided practice of written expression is the use of sentence starters (Hall & Vue, 2012). Sentence starters (aka sentences openers) are considered a specific type of prompt (Weinberger et al., 2005). The literature on the efficacy of prompts is inconclusive in that some findings are positive while others have mixed effects (Morris et al., 2010). A potential factor on whether or not a prompt produces better collaboration is the balance between structure and flexibility (Ge & Land, 2004; Morris et al., 2010). This is because too much structure may in fact impede collaboration (Morris et al., 2010). There are considered to be three categories of prompts: procedural, reflection, and elaboration (Ge & Land, 2004). Sentence starters have been used in emotional journaling to help identify learner dispositions

Culture and Social Considerations

It is well known that there are cultural differences in determining the appropriateness of emotional communication (Bagozzi et al., 1999; Shao et al., 2014). The frame of valence has been examined in cross cultural settings and found that the dimension of valence appear to be universal in the sense that people across cultures can identify emotions as either positive or negative (Russell, 1983, 1991). There are also known display rules that we in turn use as social cues for what emotion is appropriate in the current context (Malatesta & Haviland, 1982).

Emotions in Learning

The idea of contextualizing emotions in education has led to a variety of strategies that focus on examining discrete emotions that are relevant to learning (D’Mello et al., 2014; Pardos et al., 2013; Peterson et al., 2015). For example, Pekrun, Frenzel, & Goetz (2007) outlined that for education the most relevant emotions are the set of discrete emotions, called achievement emotions, which are tied directly to achievement activities or achievement outcomes. However, proponents for core affect outline that the dimensional approach is not limiting in terms of narrowing the potential emotional communication to discrete emotions (Barrett, 2006). It is important to note that while literature can, at times, be organized in terms of positive and negative emotions, it is believed that contrasting emotions in learning is favored over a hedonistic perspective (Fiedler & Beier, 2014). However, there is utility in thinking about the positive and negative emotions in the context of Piaget’s theory of constructivism as positive emotions align well with the concept of assimilation while negative emotions align well with accommodation (Fiedler & Beier, 2014). Another dimension of valence, mixed

emotions, has received some recent conceptual and empirical attention (Barford & Smillie, 2016; Berrios & Totterdell, 2013; Carrera & Oceja, 2017; Hui et al., 2009).

Self and Socially Regulated Learning

In online learning it has been pointed out that there is typically more autonomy for the students making it an interesting context to study self-regulated learning (Rienties et al., 2012). However, some have made the observation (Järvelä, 2014) that self-regulated learning was established in a context that has more of a focus on the individual differences (Zimmerman & Martinez-pons, 1990) while online collaborative environments include more social regulation sparking an extension of SRL to frame research in terms of Self and Socially Regulated Learning (SSRL) (Järvelä, 2014). Emotions have been considered to be essential for self-regulated learning (Op ' et al., 2007) and at the same time have been considered something that are socially regulated (Reeck et al., 2016). Emotions have also been acknowledged as an underexplored facet of SSRL (Järvelä, 2014; Järvelä et al., 2013).

Gaps in the Research

One controversial universal theory on emotions is core affect, which categorized emotions on the dimensions of valence, arousal, and sometimes control (Russell, 1980; Russell & Barrett, 1999b; Yik et al., n.d.). While it is a controversial theory it is very influential as the dimensions are frequently used to organize self-report of emotional response (Dan-Glauser & Scherer, 2011; Västfjäll & Gärling, 2007). What is not understood is if this categorization strategy could potentially provide an appropriate balance of structure and flexibility for sentence starters. The goal of this study is to examine an intervention of using emotional versus cognitive framing sentence starters on online communication to enable clarifying the emotional aspect of a written statement. The sentence stems will encourage people to lead their statements with an emotional categorization preface (i.e. "I had a [emotional category] reaction because..."). The categories used in the sentence stems will be: Positive, Negative, Neutral, and Mixed (e.g. "I had a positive reaction because..."). This could potentially improve communication by reducing the amount of ambiguity in a discussion – one of the limits to our interpretation of the emotional aspect of written expression (Barchard et al., 2013; Wilson et al., 2005).

5. METHODOLOGY

This study is a replication of a pilot (HREC/2016/2388/Hillaire/1). This study will take place in a computer lab setting at two sites: Maastricht University & University of Surrey. Participants will work online in groups of five on one of three collaborative tasks (which are attached to this application). The tasks will require participants to evaluate and discuss the open education data set provided online by the World Bank EdStats Dashboard (<http://datatopics.worldbank.org/education/>). The collaboration will occur within the Udio software system (described below), with participants working exclusively within an asynchronous service. This research project will incorporate a

randomised control trial method that supports three conditions and a control. In condition 1 the emotional sentence frame will be used “I had a [category of emotion] reaction to [important point] because...”.

Site 1 - Maastricht:

N=300	N=300	N=300	N=300
Control Group	Condition 1	Condition 2	Condition 3

Site 2 - Surrey:

N=70	N=70
Control Group	Condition 3

Given that the communication strategy is to use a categorization that is considered to be universal, an attempt will be made when creating groups of 5 participants to have the composition be cross-cultural.

In the control group participants will work on the group tasks as previously described by Jenna Mittelmeier (HREC/2360).

In condition 1, participants will be asked to use the following self-report of emotional reaction during the group discussion:

Self-report their emotional reaction to the group work selecting between 0 and 12 words to describe their reaction.

In condition 2, participants will be asked use of the following prompts 2 times during the group discussion

- I had a negative reaction to ...
- I had a positive reaction to ...
- I had a mixed reaction to ...
- I had a neutral reaction to ...

In condition 3, participants will be asked to use both the self-report and the sentence starters described in condition 1 and condition 2.

Participants will also be debriefed at the end of the lab (explained in detail later in this application). A rough outline of the individual lab timelines is provided below. Tentative lab schedule for each hour (70 minutes)

5 minutes Getting settled and informed consent
5 minutes pre-test emotion expressivity questionnaire
10 minutes Stats activity
5 minutes Explanation of task
35 minutes Time for group collaboration in Udio
5 minutes Post-test MES, PANAS survey
4 minutes Wrap up and debrief

24 hours after the lab exercise there will be a 1.5 hour post lab activity that will ask participants to recall their experiences with the cognitive group outcomes of the lab activity. This post lab activity is part of the learning activity for that particular week and students are expected to participate both in the lab and post-lab activity, unless they opt for the alternative assignment (see section 7). Afterwards, they are asked to critically analyse the group output and make recommendations for further refinements of the learning outcomes. These cognitive outcomes will be shared with respective group members and provide valuable learning opportunities. As a next step, participants will be asked to recall their emotional experience, examine the emotional aspect of written expression from the group exercise, and examine the emotional aspect of the work product produced by the group. In the post activity participants will be asked to recall the lab work and be asked to label their interactions with the learning content as either positive, negative, neutral, or mixed. After making that judgement they will then be asked to log into Udio and review the chat log from the group activity. After reviewing the chat log, participants will be asked to estimate how many comments were made by the entire group together. Then estimate how many comments they personally contributed to the discussion. Then estimate how many comments they made were positive, how many were negative, how many were neutral, and how many were mixed. After providing an estimate that is greater than zero each participant will be asked to provide what they believe to be the best example(s) by copy and pasting one or two comments they made that have this emotional aspect. After analyzing the emotional aspect of their comments participants will be asked if they think their peers perceived them to have a positive, negative, neutral, or mixed response to the learning material. Finally, participants will be asked to review what was produced in the learning activity. After reviewing what the group submitted the participant will be asked if what was produced was positive, negative, neutral, or mixed. Also they will be asked to rate the outcome of their work on a scale from 1 to 7 whether or not the outcome represented critical thinking by weighing multiple evidence sources, making judgements about the evidence, and generating a defensible opinion.

A week after the lab participants will be given the opportunity to participate in an interview. The interview is structured around reviewing a document that shares the analysis of comments with the participants and opens up discussion about where the participant agrees with the predictions made by the technology and where the participant disagrees with predictions made by the technology. There will be 20

interview slots with an anticipated 1 hour duration. The interviews will take place over three days.

The methodology will be quantitative in nature exploring the relationship between emotion and cognition. The basic overarching methodology will be feature generation for text classification (Forman, 2007) which can be fed into statistical models to answer questions about the relationship between the production of categorized emotional expression in text within the context of individuals and groups as it relates to outcome measures (see Appendix for description of proposed Natural Language Processing techniques). Initial findings from the pilot study last year will be used with the data collected this year as the next steps in validating a measure for sentiment detection.

Observation and video capture (where applicable) of participant will be used to provide an additional predictor of emotional response of participants. The video data will be post processed by affective computing techniques to predict emotional response. It has been noted that including some measure of human observation is good practice in emotions studies (Mauss & Robinson, 2009). While human observers will be in the classroom video of a subset of participants will be collected to generate another trace of data that can benefit from human observation. Similarly the text generated from discussion will be coded by human reviewers to get another form of human observation incorporated into the sentiment analysis (Wilson et al., 2005) which can be used as a means of verifying that sentiment analysis is not too far from machine coding.

6. PARTICIPANTS

At the first site participants are first year university students at Maastricht University, who are taking a course with Dr Dirk Tempelaar. This study builds on previous work of Simon Knight (HREC/2014/66836/Knight/2), Jenna Mittelmeier (HREC/2360), and my pilot study Garron Hillaire (HREC/ HREC/2016/2388/Hillaire/1). Participants in his module will be sent an email by their Dr Tempelaar to participate in a computer laboratory study session, which will be built into the course schedule (see next section for details). This classroom is highly diverse, with international students present from countries around Europe and the world. Thus, it is expected that participants will be come from a wide range of backgrounds, including domestic Dutch students, European students, and students from outside of Europe. Altogether around 1200 participants are expected. All procedures and materials for this study were negotiated with our collaborator and gatekeeper at Maastricht University who has managed the necessary approvals at the host institution (Dr Dirk Tempelaar).

At the second site participants are masters students in an organizational behavior course at the University of Surrey, who are taking a course with Dr YingFei Heliot.

This study builds on previous work of Jenna Mittelmeier (HREC/2360), and my pilot study Garron Hillaire (HREC/ HREC/2016/2388/Hillaire/1). Participants in his module will be sent an email by their Dr Heliot to participate in a computer laboratory study session, which will be built into the course schedule (see next section for details). This classroom is highly diverse, with international students present from countries around Europe and the world. Thus, it is expected that participants will be come from a wide range of backgrounds. Altogether around 140 participants are expected. All procedures and materials for this study were negotiated with our collaborator and gatekeeper at University of Surrey who has managed the necessary approvals at the host institution (Dr YingFei Heliot).

7. RECRUITMENT PROCEDURES

This study is built into the schedule of an introductory statistics class. Participants can opt to take part in the study, or to complete a separate course-related task of equal time and effort. Thus, participants are not required to attend the study or participate in the lab session. They may also withdraw and complete the alternative task at any time. It is worth noting that participants will not receive grades or marks for study participation or the alternative task (i.e. they will not be penalized for minimal effort or completion). Therefore, the study will not directly impact participant's grade or courses. Potential participants will receive an email from their classroom teacher one week prior to the lab, which will include our study information sheet (attached). They will be made aware at this time of the alternative assignment, as well as the notion that they are not required to attend and will not be marked for this activity.

8. CONSENT

Participants in this study will be prompted to give informed consent at the start of the lab activity. They will first be provided with a verbal briefing upon arrival, accompanied by a full online consent form (attached to this application). Participants will give consent by ticking an on-screen box, logging into the Udio system (using a unique log-in populated for individual participants), and submitting their unique ID number. In doing so, the participants will also consent to Maastricht University sharing demographic data about them with the Open University research team. Participants will be unable to progress into the activity without providing informed consent.

9. LOCATION(S) OF DATA COLLECTION

This study will take place at the Maastricht University in Maastricht, Netherlands, and University of Surrey in Guildford, England with participants in a computer lab using the Udio software system (explained below). Servers for the software system will be hosted securely by the Institute of Educational Technology on a Postgres system within the UK, however, meaning data collection from the server will take place on the OU campus. No Udio data will be stored on any servers or computers at Maastricht University or University of Surrey as browser cache will be cleared after each lab session. Demographic data about participants who agree to take part in this study will be collected from the host university (including age, gender, nation of origin, nation of citizenship). This will be sent from the classroom teacher (Dr Dirk Tempelaar & Dr YingFei Heliot) by an encrypted, password protected file via email. Participant identities in this file will be anonymised through the use of unique identification numbers in line with the Data Protection Act's definition of 'anonymized.'

Udio Description

Udio is a literacy platform developed by CAST.org, funded by the Office of Special Education Programs from the U.S. Department of Education (<http://cet.cast.org/udio/>). The platform was intended to create a curated core set of high interest reading material supplied by content partners in order to provide literacy supports for all students, including students with disabilities. While the platform was originally developed with the middle school population in mind, the Udio platform itself has a great potential to be used as a more general research platform. In partnership with CAST, researchers at the OU will explore the potential benefits of Udio as a research platform in a higher education context. The Udio platform functions within a web browser, and it is not a standalone software. Any material can be provided on the Udio platform as long as it can be formatted into an XML structure, making it possible to take most content that was developed for the internet and create a version of that material within Udio. When content is reformatted into Udio, it is referred to as an 'article.' Articles can be either single page or have formatting like chapter books with a table of contents. For a single page article there are supports on the page. The set of support features include:

- * a discussion feature that is displayed side by side with the content as well
- * a reading comprehension check
- * the ability to collection snippets of text or images from articles in Udio
- * the ability to create and publish a project based on material in Udio

For the purposes of this study, the assignments for the class will be provided as a single page article in Udio with all features available on the page of the assignment. The supports we will focus on using are the discussion supports as well as the project feature. During the lab Upon arrival to the lab, participants will be given an introduction to the activity and be prompted to provide informed consent (further described in a later section). Participants will then be asked to log into the Udio

system on a web browser using a unique log in. Inside the system, they will be asked to read their activity instructions and begin the group work activity using the discussion feature. The Udio platform will be hosted on OU servers and this installation will be used for this study.

10. SCHEDULE

This study will take place October 9 – 13, 2017 at Maastricht University.
This study will take place October 17, 2017 at University of Surrey.
Interviews will take place October 18-20, 2017 at Maastricht University.
The data will be analysed and written up for the primary investigator's PhD as part of study 2 by December 2017.

Key Ethics considerations

11. PUBLISHED ETHICS AND LEGAL GUIDELINES TO BE FOLLOWED

BERA

12. DATA PROTECTION AND INFORMATION SECURITY

This study has been registered with the University's Data Protection Coordinator. All data related to this study will be saved in a password-protected file on a university computer. All personally identifying information (such as name and Participant ID) will be removed, but unique identifier numbers will be used in line with the Data Protection Act's definition of 'anonymized.' Destruction of the data will occur at the earliest in October 2018 (or the end of this PhD project) and at the latest by October 2020, in order to allow for journal article publication and dissemination of findings.

13. RESEARCH DATA MANAGEMENT

The Udio Platform will collect log data stored on a Postgres database. During the study, all clickstream data will be logged around what materials were accessed, as well as traces of support utilization including discussion comments. At the conclusion of the study there will be a timeframe of 4 weeks where the data will remain on the server, allowing for data pre-processing using the database to format and extract data into a format ideal for conducting statistical analysis. During this month, some locally-hosted tools will be used to access data from the database for exploratory analysis purposes. Once that data have been prepared for statistical analysis, a backup of the database will be created for archival purposes to ensure that the research is reproducible. This archive will be stored in a secure manner on OU servers and maintained through 2020. Once archived, the database will only be used in the event that additional information is needed for retrieval to conduct a secondary analysis. Each of the investigators on the project will pull a pre-processed set of data from the database prior to archiving the database. The data will be to explicitly support the research questions outlined in the studies. Each of the researchers will take their data files and manage them appropriately in terms of keeping the files on OU machines and following OU data security policies. These independent files will for analysis purposes will be retained by research staff through 2020.

14. DECEPTION

None

15. RISK OF HARM

None

16. DEBRIEFING

A debriefing will be given verbally to participants at the end of the lab activity. In addition, students will receive cognitive feedback from their peers on the group outcomes and feedback from the post-lab activities, which will be a useful learning experience. A summary of the research findings will be compiled at the end of the analysis phase of this study and a copy of this report will be shared with participants via email, as well as with the Business & Economics department at Maastricht University. This report will contain no identifying information about participants. Participants will also be informed that they may contact the PI with any questions after the lab, or if they would like to withdraw part or all of their data up to 90 days afterwards. They will be given a copy of the study information sheet for their personal records, which includes contact details of the research team.

Project Management

17. RESEARCH ORGANISATION AND FUNDING

This study is part of the Leverhulme Open World Learning PhD programme.

18. OTHER PROJECT-RELATED RISKS

None

19. BENEFITS AND KNOWLEDGE TRANSFER

This research aims at better understanding of how to support emotional content to communication in online collaboration as well as how that relates to learning measurements and outcomes. There are practitioner implications as the sentence stem intervention is based on effective teacher practice methods. There are analytics implications as the use of sentiment analysis to detect emotional aspect of written communication will be examined. Finally there are research implications as the role of emotion expression in written communication will be examined in a group work context.

20. DISSEMINATING AND PUBLISHING RESEARCH OUTCOMES

The findings of this study will also be compiled for publication in the Computers & Human Behavior. Conference articles will also be written for Learning at Scale, ICLS, and EDM. All publications will be available on ORO and will be circulated via social media.

21. DECLARATION

I declare that the research will conform to the above protocol and that any significant changes or new ethics issues will be raised with the HREC before they are implemented.

I declare that I have read and will adhere to the following two OU documents:

- [OU Code Of Practice For Research and at the Open University](#)
- [OU Ethics Principles for Research involving Human Participants](#)

<http://www.open.ac.uk/research/ethics/index.shtml>)

To meet internal governance and highlight OU research, the titles of all projects considered by the HREC (whether by HREC checklist or proforma), will be added to the Research Ethics website - <http://www.open.ac.uk/research/ethics/human-research>. If you would prefer for your title **not** to be made public, or have any queries, please email the HREC Secretary on Research-REC-Review@open.ac.uk.

	Garron Hillaire
Name:	_____
	IET
Unit/Faculty:	_____
	07493079493
Telephone	_____
	Garron.hillaire@open.ac.uk
E-mail	_____
	Garron Hillaire
Signature(s) (this can be the typed name(s) of investigator(s) if an electronic copy is submitted (which is preferred))	_____
	October 4 th , 2017
Date:	_____

End of project final report

Once your research has been completed you will need to complete and submit a final report to the HREC. A copy of the template can be found on the Research Ethics website at <http://www.open.ac.uk/research/ethics/human-research/human-research-ethics-full-review-process-and-proforma#final-report>.

May 2018

Proposed date for final report: _____

APPENDIX

In order to generate features about text Natural language processing (NLP) is the application of computational methods to process written expression which includes the sentiment analysis or opinion mining. Sentiment analysis attempts to identify written expresses as positive or negative (Pang & Lee, 2006). Sentiment has a distinction from emotion in that emotions can be free floating while sentiment typically has a target object (Munezero et al., 2014).

Sentiment analysis has been used to categorize the emotional aspects of written expression, sentiment, mood, and at times model discrete emotions (Roan et al., 2009). The approach has a variety of methods including lexical approaches and machine learning approaches (Medhat et al., 2014). Lexical approaches are considered to be the simplest approach. In lexical approaches dimensions like valence (positive to negative) are computing using word substitution to score sentences by averaging the score of the words as they are previously ranked in a dictionary on the dimension of valence (Pang & Lee, 2006). Dialectical emotional complexity will be identified by comparing collocates of words with opposing values based on the dictionary substitution method (Grossmann et al., 2015). The approach can become more complex with methods that use machine learning classifiers (Medhat et al., 2014).

APPENDIX 9 – PERMISSION FORM FOR 2017

Thank you for participating in this lab assignment.

Activity data for this assignment will be collected for research purposes. Below is information about this joint study.

Study title: Investigating tools to support emotion communication during cross-cultural collaborative group work

What is the purpose of the study?

We are inviting you to take part in a study evaluating how diverse groups work together and if different types of academic content play a role in task behaviour and discourse as well as the potential for sentence starters to clarify communication.

Why have I been approached?

For the purposes of the study we need to recruit a number of groups of students studying in a higher education institution.

Do I have to take part?

No. Participation is entirely voluntary. If you change your mind about taking part in the study you can withdraw at any point during this session and at any point up to 90 days after the session. If you decide to withdraw, all your data will be destroyed and will not be used in the study. There are no consequences to deciding that you no longer wish to participate in this laboratory activity. There is an alternative assignment from your teacher for those who do not wish to participate.

What happens during the study?

You will be able to complete this study from a quiet location in a computer lab where you won't be disturbed. You will be working in small groups of five participants to complete a problem based learning

task. You will be working in a different location to your group members. Altogether, this lab activity will take 70 minutes, and you will be asked to complete a post-activity at home which will take approximately two hours.

The study will involve collaboration via an online instant messaging system in your web browser with group members for no longer than 45 minutes. During this time, you will be asked to work on a sampling activity, review and reflect on educational data available from web resources. During the post-activity task, you will be asked to reflect on your group's collaboration process and soft skills of working with diverse peers. Some participants will be selected to use sentence starters that support emotional elements of communication.

To complete this task, you will be asked to log in a website called Udio. While you are logged in, the website will collect data of your instant messaging conversation. In collaboration with Dr Dirk Tempelaar we will collate anonymised demographic for analysis about how students learn together, including your gender, age and nation of origin. Some participants will be video recorded. The reason for video capture is to look for emotional reactions during discussions by observing the video, using artificial intelligence that interprets facial expression, and artificial intelligence that predicts heart rates based on techniques like motion capture. If you are located at a desk where a web camera is recording you will be notified at the start of the lab and asked if you would like the cameras turned away during the activity to avoid being recorded.

What are the possible disadvantages and risks of taking part?

We do not anticipate any risks associated with participation in this study.

What are the possible benefits of taking part?

You will gain an insight into how a psychology research project is conducted and what it is like to be a participant in such a study. The tasks are relevant to all students as they are about the kind of transferable skills around collaboration, inferring about data and finding information, that all graduates should have. You will also receive individualised feedback about your contributions to your group, including relevant soft skills for collaboration.

Will my taking part in this study be kept confidential?

Yes, no personally identifying information will be shared.

What will happen to the results of the research study?

You will be given a personal login for the Udio system. Identifying information will be kept passworded, and will not be associated with the anonymous data collected.

This research forms part of Garron Hillaire's doctoral research at the Open University supervised by Prof. Bart Rienties, Prof. Mark Fenton-O'Creevy, and Prof. Zdenek Zdrahal. All data will be available to Dr Dirk Tempelaar in his capacity as course leader.

Deanonimised data will not be shared other than within the OU supervisory and external examiner team, except where we are legally bound to do so. Pseudonyms will be used in reporting, and any identifying information mentioned in the instant messenger logs will be redacted.

The data will be kept in full for the duration of the investigator's PhD research or until December 2018 (whichever is later). Data stored will be kept in a password protected file in accordance with the Data Protection Act.

Who is organising and funding the research?

The research is organised by Garron Hillaire a research student at the Open University's Institute of Educational Technology. This work is funded by the Open University and the Leverhulme Trust.

Who has reviewed the study?

The Open University Ethics Committee has reviewed and approved this study.

Contact for Further Information

Garron Hillaire
Institute of Educational Technology, Walton Hall, Milton Keynes, MK7 6AA
Email: Garron.hillaire@open.ac.uk

Agreement to participate

[Tick box here]

I understand that checking this box constitutes a legal signature confirming that I acknowledge and agree to the above terms.

[Finish button]

APPENDIX 10 – INTERVIEW PROTOCOL FOR 2017

Total Time 60 minutes

The interviews will take place in an office with two seats in front of a computer. The interview will start with a couple of warm up questions then we will use a document on the computer to review some data together. The interview will conclude with a couple of open ended questions. If at any time a participant asks to stop or looks visibly upset the exit protocol will be used to end the interview.

Introduction: (5 Minutes)

Hello and thank you for participating in this interview where we are providing a detailed report of how your participation in the lab was analysed in term of emotion. My name is Garron Hillaire and I am the primary investigator in this study. Is it ok if I record this interview? [if yes, turn on recorder]

Warm up: (10 Minutes)

Before we get started I wanted to ask a few questions about your experience in the lab last week.

1. Did you enjoy the lab activity where you discussed real world statistics to make a funding decision?
2. Do you have any suggestions on what would make the activity better?

Review Comments: (10 Minutes)

Great thanks. Just to recap we are going to look at the emotions expressed in your chat messages. First I would like to review some comments with you and ask what emotion you. The messages we will review were randomly selected from your group discussion during the lab. If any of the messages make you uncomfortable we do not need to discuss them. In fact, If this process becomes uncomfortable for any reason at all you please let me know and we can stop the interview. We are going to look at messages and I would like you to select between positive, negative, neutral, and mixed to classify the message. If there is anything tricky about the message we will write that in the notes

Comment	Valence	Notes
548) dankewell	Positive Negative Neutral Mixed	
549) great	Positive Negative Neutral Mixed	

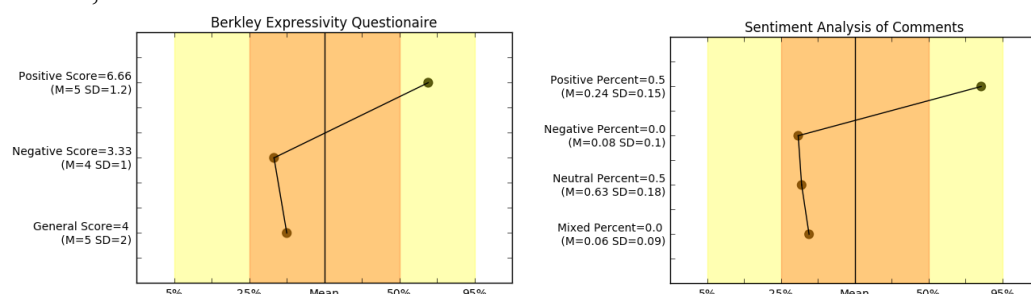
Review Sentiment Analysis: (10 Minutes)

Researchers on emotions in learning say that there are benefits to both positive and negative emotion so do not consider predictions of negative expression as bad and positive as good. Rather consider that learning involves both positive and negative emotion. Again the messages were randomly selected and the predictions made by technology are not 100% accurate. I will explain how to read the first message and we will compare your prediction to see if you agree with the technology and after reviewing the explanation together you will let me know if the technology changes your mind after reading the explanation.

Comment	Analysis	Details	Valence	Agreement	Change
548)	dankewell[0]	[[Sentence=-1,1=word max, 1-5]]	Neutral	Yes No	Yes No
549)	great[2]	[[Sentence=-1,3=word max, 1-5]]	Positive	Yes No	Yes No

Look at Aggregate Scores: (10 Minutes)

Here are two charts the first is your scores on emotional expressivity and the second is the comments you made during the group chat in categories of positive, negative, neutral, and mixed.



Closing Questions: (15 Minutes)

1. Do you think sentiment analysis of your comments is useful when provided for individual sentences?
2. Do you think sentiment analysis of your comments is useful when provided to you in aggregate form?
3. Is there anything you were hoping to see that you did not see when looking at the sentiment analysis of your comments?

Exit Protocol

1. If a participant explicitly asks to stop respond with:
 - a. I completely understand. The interview is over. Are you ok? All records of the interview will be deleted. I am sorry about this. Can I get you a glass of water (or tissue if appropriate). You are free to stay here and regroup. Would you like some privacy for a few minutes? Would you like to talk about something else? You are free to leave and again I apologize for this.
2. Participant is visibly upset

HUMAN RESEARCH ETHICS COMMITTEE (HREC)
AMENDMENT FORM



- a. Are you doing ok? We can stop the interview if you like? Can I get you a glass of water (or tissue if appropriate). You are free to stay here and regroup. Would you like some privacy for a few minutes? Would you like to talk about something else? You are free to leave and again I apologize for this.

APPENDIX 11 - ADDENDUM

Please return the completed form to Research-rec-review@open.ac.uk. Amendments will be reviewed by HREC and you will be notified of the outcome within 7 working days.

SECTION 1. PROJECT DETAILS	
Project Title:	Maastricht RCT of Cognitive & Emotion Sentence Starters
Applicant's Name:	Garron Hillaire
HREC Reference:	HREC/2388 & Previous addendum "Replication of Maastricht RCT of Emotion Sentence Starters by Garron Hillaire (Previous approval HREC/2388)"

SECTION 2. SUMMARY OF AMENDMENTS TO THE PROJECT
<p>Please note any amendments to the study should be outlined and highlighted on the original application form and resubmitted with this amendment summary form.</p> <p><i>Briefly summarise the main changes proposed in this amendment e.g. a change to the research methodology, inclusion of a new group of participants, a change in location or an addition to the content of the study in some way e.g. a new questionnaire for participants. Explain the purpose of the changes and their significance for the study.</i></p>
Changes to the project
The changes I propose to the project are changes to the team and end date
Changes to the research team
1. Corrections from my thesis defence requested

- a. "The candidate is strongly advised to run a comparison between the annotations provided by participants in his study and those provided by sources external to his study (e.g. the crowdsourcing website Amazon Mechanical Turk). The aim of this would be to provide evidence that these factors (internality/externality) influence the outcomes of sentiment analysis."

To address this concern I would like to use the Mechanical Turk service to get labels of positive/negative/neutral/mixed for anonymized chat message data from homogenous raters (7 raters from the Netherlands) and diverse raters (20 raters from a variety of countries). In addition, I have coded this data myself, and would like to share anonymized data with Jenna Mittelmeier and Francisco "Paco" Iniesto to also rate anonymized chat messages.

Changes to the proposed study end date

2. As the correction and submission delays pushed past the original project deadline I would also like to amend the study end date to October 2021 as a measure to have time for further correction in the event my revisions require further work or a requested extension to continue revising my thesis is granted

SECTION 3. ETHICAL CONSIDERATIONS & DATA PROTECTION

Are any ethical issues raised by the changes? If so, how will they be addressed?

The main ethical issue raised here is sharing chat messages from students with external parties: Mechanical Turk, Jenna Mittelmeier & Paco Iniesto. To mitigate the risk I will only share anonymized data where student names and references to their country of origin in the chat messages will be redacted.

Are any data issues raised by the changes? If so, how will they be addressed?

*Please consider the data processing arrangements for **any** new data that will be collected, shared or stored as part of this amendment. Any changes to existing data collection methods/storage arrangements should also be summarised here.*

The main change is sharing the text messages (anonymized) with coders from mechanical turk, Jenna Mittelmeier & Paco Iniesto, to identify if the messages are positive/negative/neutral/mixed.

SECTION 4. SUPPORTING DOCUMENTS

Please include any documents related to the proposed amendment as separate attachments and indicate which documents you are including below e.g. a consent form and participant information sheet for a new group of participants.

Where there is more than one group of participants, please provide separate consent forms and participant information sheets for each group.

Consent form and Participant information sheet – for each participant group	<input checked="" type="checkbox"/>
Questionnaire (for online surveys please include either a Word version of the questions or a link to the survey online)	<input type="checkbox"/>
Email or letter from the organisation agreeing that the research can take place	<input type="checkbox"/>
Draft bid or project outline	<input type="checkbox"/>
Publicity leaflet	<input type="checkbox"/>
Data Management Plan	<input type="checkbox"/>
Other	<input type="checkbox"/>

Please continue to next page

SECTION 5. DECLARATION

I declare that:

- I understand that I must not implement any amendments to a previously approved study until I have received approval from HREC
- The research will conform to the protocol outlined above and I will inform HREC of any subsequent amendments to the protocol of this study and have these agreed before they are implemented
- I have read and will adhere to the following OU policies:
 - OU Code of Practice for Research – (see the [Research Ethics Guidelines page](#))
 - OU Ethics Principles for Research with Human Participants – (see the [Research Ethics Guidelines page](#))

Principal Investigator (Name):

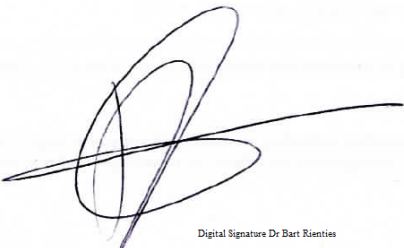
Garron Hillaire

Principal Investigator (Signature): Garron Hillaire

Date Sept 2nd 2020

FOR STUDENTS ONLY:

*Please note that this amendment cannot be processed without your **OU supervisor's** signature and supporting comments, which should be provided below.*

Postgraduate research degree	MPhil/PhD <input checked="" type="checkbox"/>	EdD <input type="checkbox"/>	DHSC <input type="checkbox"/>
Student Personal identifier	E5780548		
OU Supervisor's name	Bart Rienties		
OU Supervisor's email address	Bart.rienties@open.ac.uk		
OU Supervisor's electronic signature	 Digital Signature Dr Bart Rienties		

OU Supervisor's supporting comments: The validation of the anonymized student comments by a third party would allow Garron to address the concerns by the external examiner that the coding of the written chat data is appropriate. As the student chat data will be anonymized by Garron, and the chats were focused on tasks whereby students worked in an anonymous/blinded design on a 40 minute computer task, there is limited to no risk that any personal data will be shared to a third party. Furthermore, as in Mechanical Turk raters will only see a random selection of a small number of sentences it is unlikely that any sensitive information can be linked together.